Tehnici de procesare a semnalelor audio
Preprocesarea semnalelor audio

EUGEN LUPU, CORNELIU RUSU

Eugen.Lupu@com.utcluj.ro, Corneliu.Rusu@bel.utcluj.ro

Cuprins

- Caracteristici subiective și obiective ale semnalului audio
- Caracteristici statistice ale semnalului audio
- Analiza semnalului audio
- Analiza în timp a semnalului audio
- Metode de detecție vorbire/liniște
- Metode de detecție a frecvenței fundamentale
- Analiza în frecvență a semnalului audio

Caracteristici subiective și obiective ale semnalului audio

Pentru definirea sunetelor în percepția auditivă s-au stabilit corespondențe ale caracterelor semnalelor acustice:

- înălțimea - care este corespondentul subiectiv al frecvenței semnalului;
- intensitatea (tăria) - care este echivalentul subiectiv al intensității acustice, fiind dată de amplitudinea undei acustice;
- timbrul sonor - corespunde altor caracteristici care conferă individualitate sunetelor, pe baza lor urechea putând recunoaște sunetele chiar dacă prezintă o serie de caracteristici comune. În general timbrul se referă la compoziția spectrală la care se adaugă și evoluția intensității în perioada tranzitorie.

Caracteristici statistice ale semnalului audio

- **densitatea de amplitudine** - reprezintă distribuția statistică a amplitudinii eșantioanelor, ceea ce permite scalarea corectă a unui semnal vocal aplicat la intrarea unui convertor analog/numeric;

- **valoarea medie a spectrului pe timp lung** - pot exista distribuții ale spectrului specifice unor clase de semnale;

- **distribuția frecvenței fundamentale** - experimentele făcute arată că distribuția frecvenței fundamentale, atât la bărbați cât și la femei, este gaussiană (normală) cu valoarea medie de 125 Hz la bărbați și de 250 Hz la femei, având o deviație standard de 20 Hz, respectiv de 40 Hz;

Analiza semnalului audio

Analiza semnalului audio urmărește extragerea unor parametri din semnal care să fie utili aplicației respective.

- Analiza spectrală a semnalului audio pe o perioadă scurtă de timp numită și **analiza pe termen scurt**, poate furniza informații atât despre tract, cât și despre excitație.
  - Reprezentările spectrale produc un semnal bidimensional, în timp și frecvență.
  - Dimensiunea temporală se referă la poziția segmentului analizat iar dimensiunea frecvențială se referă la componentele spectrale ale segmentului.
  - Aceste reprezentări bidimensionale se pot prezenta sub forma spectogramelor (reprezentări amplitudine - timp - frecvență).

- **Analiza pe termen mediu** furnizează vectori dinamici, în care se iau în considerare câteva cadre succesive dintr-un anumit interval de timp și se generează informații despre gradientul spectral.
  - Parametrii vizează viteza de modificare a parametrilor statici (parametri delta) sau accelerația acestora (parametri dublu-delta sau delta-delta)
  - Parametrii delta se calculează prin interpolarea parametrilor statici din câteva cadre consecutive ce acoperă un interval de timp de 50-100 ms.
  - Parametrii de tip accelerație se estimează pe intervale de timp mai lungi.
  - Calculul acestor parametri dinamici poate fi văzut și ca o filtrare a traiectoriilor temporale ale parametrilor statici cu filtre FIR trece-bandă.

Analiza în timp a semnalului audio

- **Amplitudinea maximă și medie a semnalului audio** dau informații calitative despre prezența sau absența semnalului audio sau eventual dacă segmentul analizat este sonor sau nesonor:

$$M(n) = \frac{1}{N}|x(n)| * w(n), \tag{1}$$

unde $x(n)$ este semnalul de analizat, $w(n)$ este fereastra de analiză (de obicei rectangulară), iar N este numărul eșantioanelor semnalului de analizat.

Segmentele sonore (vocalele) au o amplitudine medie semnificativă, pe când segmentele nesonore (consoanele) se aseamănă cu zgomotul, având amplitudinea medie scăzută.

- **Energia medie pătratică** pentru un semnal cu valoarea medie nulă sau **energia pe timp scurt** se poate defini astfel:

$$E(n) = \frac{1}{N} \sum_{m=\infty}^{\infty} x^2(n) w^2(n-m) \tag{2}$$

Segmentele sonore (vocalele) au o energie importantă, pe când vorbirea nesonoră (consoanele) se aseamănă cu zgomotul, având energie scăzută.

- **Operatorul de energie Teager** este un mod de calcul al energiei instantanee:

$$E_i = x_i^2 - x_{i+1} x_{i-1}. \tag{3}$$

Această formă de energie instantanee se modifică atât cu amplitudinea cât și cu frecvența semnalului audio.

Operatorul Teager poate fi folosit pentru separarea zonelor de liniște într-un semnal vocal.

- **Energia medie Teager** - energia medie se poate evalua prin însumare și mediere:

$$E = \sqrt{\sum_{i=1}^{N} \omega_i^2 |X(\omega_i)|}, \tag{4}$$

unde $X(\omega_i)$ este coeficientul Fourier nenul pentru frecvența ω_i.

Zonele nevocalizate din semnalul vocal sunt mai bine puse în evidență decât în cazul energiei medii pătratice.

- **Numărul trecerilor prin zero** reprezintă o aproximare a conținutului spectral al semnalului vocal. Pentru un semnal cu N eșantioane numărul trecerilor prin zero se calculează astfel:

$$\mathrm{NTZ} = \frac{1}{2}\sum_{n=0}^{N-1}\{1 - \mathrm{sign}[x((n+1)T)]\mathrm{sign}[x(nT)]\} \tag{5}$$

Numărul trecerilor prin zero necesită o normalizare în raport cu valoarea maximă a acestuia corespunzătoare frecvenței maxime din banda semnalului și cuantizare, prin stabilirea unor praguri și clase disjuncte pentru valorile obținute.

În urma acestor prelucrări, numărul trecerilor prin zero poate fi folosit ca și caracteristică globală pentru recunoaștere.

Valorile numărului trecerilor prin zero sunt puternic dependente de

frecvența de eșantionare și de mărimea ferestrei pe care se face numărarea.

Valoarea NTZ împreună cu valoarea energiei permite stabilirea caracterului sonor/nesonor precum și prezența sau absența semnalului vocal.

De NTZ este legată inteligibilitatea vorbirii. Porțiunile sonore prezintă un număr mai redus de treceri prin zero, pe când cele nesonore sau zgomotele au valori mai mari ale numărului de treceri prin zero.

NTZ poate fi folosit la determinarea periodicității semnalului audio, adică a frecvenței sale fundamentale.

Metode de detecție vorbire/liniște

Un algoritm de detecție automată a prezenței semnalului vocal (*endpoint algorithm*) trebuie să răspundă următoarelor cerințe:

- **siguranță și robustețe**, adică să nu greșească în cazul unui raport semnal-zgomot variabil sau în cazul unui semnal vocal cu tărie variabilă;
- **acuratețe** - să nu piardă segmente vocale scurte;
- **adaptabilitate** - să se poată adapta la un zgomot de fond variabil;
- **simplitate**;
- **procesare în timp real**;

- să lucreze bine fără cunoștințe a priori despre zgomotul de fond.

- Mărimile esențiale pentru detectarea caracterului sonor, nesonor sau zgomot pentru un segment de semnal vocal, în domeniul timp, sunt **energia medie** și **numărul trecerilor prin zero**.

- Pe baza acestora se poate lua decizia pentru un segment de vorbire dacă este sonor, nesonor sau dacă avem o zonă de liniște sau de zgomot:

SEGMENT	ENERGIE	NTZ
Sonor	Mare	Mic
Nesonor	Medie /Mică	Mediu/Mare
Zgomot	≈ 0	Mare
Liniște	≈ 0	≈ 0

- Pentru a realiza o separație cât mai precisă trebuie găsite praguri adecvate

pentru energie și numărul trecerilor prin zero.

- **Energia medie** și **amplitudinea** au fost folosite la detecția automată a zonelor de liniște.

- În cazul energiei medii, determinarea zonelor de vorbire/liniște se poate face după următorul algoritm:

1. Se elimină componenta continuă din semnalul vocal pentru ca numărul trecerilor prin zero să poată fi calculat corect;

2. Se determină valoarea medie (Em) și deviația standard (Estd) pentru energie folosind primele 100 ms ale semnalului vocal, zonă în care se consideră că este prezent doar zgomotul de fond;

3. Se determină următorul prag pentru energia zgomotului de fond:

$$\mathrm{ETL} = \mathrm{Em} + 3\mathrm{Estd};$$

4. Se determină următorul prag pentru energia semnalului:

$$\mathrm{ETH} = 1.5 \cdots 3\mathrm{ETL};$$

5. La începutul semnalului vocal, se caută primul cadru (N1) în care energia depășește pragul ETH; acesta va fi prima estimare pentru primul cadru al zonei de vorbire;

6. Pornind de la acest cadru se merge înapoi până acolo unde energia scade sub nivelul pragului ETL; de la acest cadru se mai pot lua un număr de

cadre (uzual 3) pentru ca trecerea de la zona de liniște la cea de vorbire să nu se facă brusc;

7. Procesul se repetă la sfârșitul zonei de vorbire pentru a determina ultimul cadru din zona de vorbire (N2) .

- Energia medie Teager poate fi folosită la detecția automată a zonelor de vorbire-liniște într-o manieră similară. Deosebirea apare la alegerea pragurilor implicate în algoritm.

Metode de detecție a frecvenței fundamentale din semnalul audio

- Frecvența fundamentală (F0) este mărimea care se definește pentru porțiunile sonore ale semnalului vocal și reprezintă frecvența de oscilație a coardelor vocale.
- Corespondentul acustic al frecvenței fundamentale este frecvența fizică; Frecvența percepută se numește *pitch*.
- Frecvența fundamentală poate lipsi din spectrul semnalului vocal (de exemplu în vorbirea telefonică).
- Variația frecvenței fundamentale în timpul vorbirii este lentă, sub 10 Hz.

- Media și deviația standard ale frecvenței fundamentale pentru vocile masculine este de 125 Hz și respectiv 20.5 Hz, iar pentru cele vocile feminine aceste valori sunt duble.

- Pentru porțiunile de vorbire vocalizată, frecvența fundamentală este cuprinsă între 50 Hz și respectiv 500 Hz, iar pentru porțiunile nevocalizate se consideră, prin convenție, că F0 are valoarea 0.

- Un semnal vocal nu are o perioadă fundamentală pe întreaga durată: pe timpul pauzelor în vorbire sau pentru sunetele nevocalizate, perioada fundamentală lipsește.

- Există mai multe modalități de calcul al frecvenței fundamentale:

  – în domeniul timp;
  – în domeniul frecvență;

– prin metode timp-frecvență.

- În domeniul timp majoritatea metodelor se bazează pe funcția de autocorelație.

- În continuare vor fi prezentate metoda autocorelației, metoda funcției diferență de amplitudine medie și metoda cu limitare centrală.

Metoda autocorelației

Funcția de autocorelație este definită astfel:

$$R(k,n) = \sum_{m=-\infty}^{\infty} x(m+n)w(m)x(n+m+k)w(m+k). \qquad (6)$$

- Pentru un semnal periodic funcția de autocorelație prezintă o periodicitate cu aceeași perioadă ca și cea a semnalului inițial.
- Funcția de autocorelație prezintă maxime distincte la intervale de timp cu durata de 1/F0.
- Se caută determinarea poziției acestor maxime ce apar în funcția

de autocorelație, operație denumită *peak detection* în literatura de specialitate.

- Distanța dintre aceste maxime dă perioada fundamentală.

- În cazul funcției de autocorelație, aceste maxime sunt descrescătoare în amplitudine, ceea ce ușurează detectarea acestora.

- Operația de detectare a vârfurilor nu este simplă și influențează mult performanțele metodei.

- Dacă se iau în considerare anumite ipoteze în legătură cu aceste vârfuri (de exemplu că sunt descrescătoare în amplitudine), operația devine mai simplă, dar susceptibilă la rezultate eronate.

- Lungimea ferestrei trebuie să îndeplinească anumite cerințe pentru a obține rezultate valide:
  - Pe de o parte lungimea ar trebui să fie mică datorită variabilității semnalului audio;
  - Pe de altă parte, pentru a putea obține informația de periodicitate, este nevoie ca fereastra să cuprindă cel puțin două perioade ale formei de undă.

- Pentru obținerea unor rezultate cât mai bune se folosește o prelucrare prealabilă a semnalului pentru a îndepărta unele efecte ale tractului vocal și unele neregularități ale semnalului, datorate zgomotului sau caracterului cvasiperiodic al semnalului vocal.

Metoda funcției diferență de amplitudine medie

Funcția diferență de amplitudine medie (AMDF: Average Magnitude Difference Function) are expresia:

$$D(n) = \frac{1}{N} \sum_{k=0}^{N-1} |s_k - s_{k-n}|. \tag{7}$$

- Funcția $D(n)$ prezintă minime la intervale de timp de dată 1/F0, minime care, în general, descresc în amplitudine.

- Valorile funcției $D(n)$ sunt mai ușor de calculat decât valorile funcției de autocorelație deoarece nu implică multiplicări, ceea ce o face să se preteze la implementări hardware.

- Apar situații când aceste minime nu sunt descrescătoare în amplitudine, ceea ce conduce la obținerea unor valori duble ale perioadei.

- Aceste situații apar în cazul când lungimea segmentului vocal luat în considerație nu este suficient de mare, chiar dacă acoperă 2-3 perioade de semnal.

- Și această metodă furnizează rezultate eronate în zonele de trecere, la începutul și sfârșitul unui segment vocal.

Metoda cu limitare centrală

Metoda cu limitare centrală (*CLIP: Center Clipping Pitch Detector*) folosește un prag $\mathrm{C_L}$ și se rețin numai acele elemente din semnalul vocal a căror valoare absolută depășește pragul $|\mathrm{C_L}|$.

Pentru valorile rămase se procedează astfel:

- Se adaugă valoarea $\mathrm{C_L}$ la cele pozitive;
- Se scade valoarea $\mathrm{C_L}$la cele negative;
- Se atribuie valoarea maximă pozitivă, respectiv valoarea maximă negativă (limitare infinită);

- În continuare se calculează funcția de autocorelație.

  Aceasta prezintă maxime la intervale de timp direct corelate cu perioada asociată frecvenței fundamentale.

- Metoda este sensibilă la valoarea de prag folosită.

- Se poate folosi pentru $\mathrm{C_L}$ valoarea medie sau valoarea maximă a semnalului pe segmentul considerat, eventual înmulțită cu un coeficient subunitar.

- Scăderea pragului apropie forma funcției de cea de autocorelație obișnuită.

- În cazul folosirii unei valori crescute pentru prag (60%), forma funcției de autocorelație este mai simplă.

Analiza în frecvență a semnalului audio

- Metodele de analiză în frecvență consideră semnalul ca o superpoziție de sinusoide (analiza Fourier) și permit extragerea unor parametri altfel greu de pus în evidență în domeniul temporal.

- Întrucât semnalul audio se poate considera cvasistaționar pe perioade scurte de timp se folosește conceptul de analiză pe timp scurt, folosind cadre de N eșantioane succesive, rezultând un set de parametri care caracterizează segmentul analizat din punct de vedere spectral.

- Trunchierea unei părți din semnalul vocal poate avea un efect nedorit asupra rezultatului, motiv pentru care este importantă alegerea optimă a *lungimii secvenței* și a unei funcții de ponderare *fereastra*, care să minimizeze efectele de la capetele cadrului prin trunchierea secvenței de eșantioane.

Analiza prin bancuri de filtre digitale

- Scopul acestui tip de analiză este de a da o măsură a energiei semnalului vocal într-o anumită bandă de frecvență.
- Semnalul vocal este trecut pentru început printr-un set de filtre trece-bandă, după care semnalele $s_i(n)$ sunt trecute prin neliniarități de tipul redresare mono sau dublă alternanță.
- Rolul acestor neliniarități este acela de a deplasa spectrul semnalelor de intrare către zona frecvențelor joase și de a crea imagini în zona frecvențelor înalte.
- Filtrul trece-jos are rolul de a elimina imaginile din domeniul frecvențelor înalte.

- Ieșirea filtrului trece-jos $u_i(n)$ fiind o estimare a energiei semnalului vocal în banda de frecvență i.

- Filtrul trece-jos poate fi un simplu integrator, dar și un filtru IIR (Infinite Impuse Response) cum ar fi un filtru Bessel, dacă se dorește eliminarea unor neliniarități de fază.

- Blocul de subeșantionare permite reeșantionarea semnalelor $u_i(n)$ la o rată de 40-60 Hz, lucru ce permite o reducere a volumului de date pentru reprezentare.

- Ultimul bloc permite compresia domeniului dinamic al semnalului prin utilizarea unei scheme de compresie în amplitudine:

  – reprezentarea logaritmică;
  – reprezentarea μ–law.

- Alegerea numărului filtrelor digitale trebuie să țină cont de mai multe considerente:
  - Un număr prea mic nu permite bancului de filtre să acopere în întregime spectrul semnalului audio;
  - Un număr prea mare de filtre poate însemna o lățime de bandă prea mică și ca urmare în anumite benzi de frecvențe energia semnalului vocal va fi prea scăzută.

- Pentru bancuri de filtre uniforme, de regulă numărul filtrelor variază între 8 și 32.

Analiza prin transformata Fourier rapidă

- Transformata Fourier rapidă permite calculul eficient al transformatei Fourier discrete:

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp\left(\frac{-j2\pi kn}{N}\right), \quad k = \overline{0, N-1}. \tag{8}$$

- Pentru o rezoluție în domeniul timp de $T = 1/fe$, va rezulta o rezoluție $\Omega = 1/NT$ în domeniul frecvență.

- Reprezentările frecvențiale folosesc spectrul de putere $|X(k)|$, adică neglijează informația de fază conținută în spectrul $X(k)$. Acest lucru se datorește faptului că urechea este, în mare masură, insensibilă la fază.

- O primă formă de reprezentare frecvențială este spectograma.

- Această formă de analiză a fost intens utilizată în analiza semnalului vocal precum și în alte domenii.

- Spectograma se obține prin reprezentarea spectrului de putere, de obicei pe o scară logaritmică, în funcție de frecvență, dar și de timp.

- Amplitudinea spectrului este reprezentată ca intensitate de culoare (negru sau culoare).

- De regulă, se reprezintă secvențe de spectre de putere obținute pentru segmente de date rezultate în urma unei ferestruiri, folosind tehnica ferestrei glisante cu un factor de suprapunere de 25-50%.

- Spectogramele pot fi de bandă îngustă (*narrowband*) sau de bandă largă (*broadband*).

- În cazul spectogramelor de bandă îngustă, fereastra de analiză este completată în întregime cu datele semnalului vocal, iar spectrul rezultat va avea o rezoluție mare în frecvență și astfel va pune în evidență armonici individuale.

- Zonele mai întunecate corespund unor componente cu frecvențe mai mari.

- Spectogramele de bandă largă folosesc doar o parte din fereastra de analiză completată cu eșantioane din semnalul vocal, restul fiind completat cu zero.

- Se obține un spectru netezit în care zonele întunecate vor fi mai largi.

- Cu cât fereastra utilă este mai mică, cu atât efectul de netezire este mai mare.

- Spectrul de putere este aproape exclusiv reprezentat pe o scară logaritmică din următoarele motive:
  - La o modificare a câștigului se păstrează forma spectrului, dar apare o deplasare în sus sau în jos a acestuia;
  - Efectele convolutive generate de exemplu de acustica camerei sau de variații între liniile telefonice, ce apar ca efecte multiplicative în spectrul liniar, vor apărea ca efecte aditive în spectrul logaritmic și ca urmare pot fi mai ușor separate;
  - Distribuția statistică a spectrului logaritmic are proprietăți convenabile pentru metodele de recunoaștere statistice.

- Înainte de calculul spectrului de putere, semnalul vocal este, de regulă, trecut printr-un filtru numeric trece-sus de preaccentuare (*preemphasis*) pentru a crește contribuția componentelor spectrale de frecvență mai ridicată.

Analiza prin predicție liniară

- Predicția liniară (*Linear Predictive Coding - LPC*) este o metodă ce își propune să modeleze optim spectrul ca un proces autoregresiv.

- Parametrii modelulului se bazează pe teoria erorii pătratice medii minime.

- Fiind dat un semnal $s(n)$, se caută să se modeleze semnalul ca o combinație liniară a eșantioanelor precedente:

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + e(n) \tag{9}$$

unde:

– p este numărul de coeficienți din model sau ordinul modelului;

- coeficienții a_k sunt coeficienții de predicție liniară;
- $e(n)$ este eroarea de predicție;

- Parametrii care caracterizează modelul sunt:
  - Decizia sonor/nesonor;
  - Frecvența fundamentală pentru segmentele sonore;
  - Coeficienții filtrului;
  - Câștigul (al componentei continue) modelului G.

- Pe durate scurte de timp (10-30 ms) acești parametri se pot considera constanți datorită caracterului cvasistaționar al semnalului audio.

- Există două metode care se folosesc pentru minimizarea energiei semnalului de eroare: metoda autocorelației și metoda covarianței, metode care diferă prin fixarea limitelor pentru $e(n)$ pentru care se face minimizarea.

- Se folosește de obicei metoda autocorelației datorită proprietăților pe care le prezintă matricea sistemului, aceasta fiind o matrice Toeplitz, care permite rezolvarea printr-o metodă iterativă și anume prin algoritmul Levinson-Durbin. De asemenea metoda autocorelației garantează stabilitatea filtrului.

- Odată determinați coeficienții LPC și câștigul modelului G, se poate determina spectrul LPC care se folosește cu succes în recunoașterea vorbirii/vorbitorului.

- Analiza LPC poate fi folosită și pentru determinarea frecvenței fundamentale folosind metoda SIFT (*Simple Inverse Filter Tracking*).

  Această metodă utilizează eroarea de predicție pentru care calculează funcția de autocorelație, funcție ce prezintă maxime locale la intervale de lungime $1/F0$.