STUDIA UNIVERSITATIS
Revistă științifică a Universității de Stat din Moldova, 2011, nr.2(42)

ANALIZA CEPSTRALĂ ÎN SISTEME DE RECUNOAŞTERE A VORBIRII

Vladimir DUBINEANSCHI

Universitatea Agrară de Stat din Moldova

Automatic Speech Recognition (ARS) has progressed considerably over the past several decades, but still has not achieved the potential imagined at its very beginning. Almost all of the existing applications of ASR systems are PC based. This publication describes the possibility of using spectral analyses in the speech recognition.

Material şi metodă

Recunoaşterea vorbirii este la ora actuală un domeniu în creştere datorită succeselor din ultimii ani în prelucrarea numerică a semnalelor, în dezvoltarea de aplicaţii cu procesoare specializate şi în rapida dezvoltare a tehnicii de calcul. Sistemele actuale de recunoaştere a vorbirii se situează deocamdată în limite restrânse ale parametrilor caracteristici şi dedicate unor aplicaţii specifice. Din punctul de vedere al dimensiunii vocabularului şi al modului de vorbire, sistemele de recunoaştere cu performanţe acceptabile se împart în trei categorii principale.

- sisteme cu vocabular mic (10-100 cuvinte);
- sisteme cu vocabular mediu şi mare şi vorbire izolată (10 000-20 000 cuvinte);
- sisteme cu vocabular mediu şi vorbire conectată sau continuă, restrictivă la un domeniu de aplicabilitate (1 000-5 000 cuvinte).

Cele mai multe sisteme realizate aparţin claselor sistemelor mici şi mijlocii cu recunoaşterea vorbirii izolate. Sistemele de recunoaştere a vorbirii continue, în marea lor majoritate, există doar în formă experimentală, în condiţii de laborator. Chiar şi sistemele utilizate în practică, cele pentru vorbirea izolată sau conectată, nu sunt destul de robuste la zgomotul mediului în care funcţionează şi la variabilitatea vorbirii. Toate sistemele dau performanţe mai bune, dacă numărul de utilizatori este mai redus şi dacă cei care folosesc sistemul sunt cei a căror voce s-a folosit pentru învăţarea sistemului. Performanţele se degradează semnificativ, dacă vorbitorii se schimbă sau dacă sistemul este folosit cu alte cuvinte decât cu cele pentru care a fost antrenat. Caracteristicile principale ale unui sistem de recunoaştere automată a vorbirii, fără a aminti parametrii şi metodele specifice prin care s-a implementat, sunt următoarele:

- dimensiunea vocabularului, adică numărul de cuvinte capabil sa le recunoască
- monolocutor sau multilocutor (aici se poate preciza şi sexul vorbitorilor)
- vorbirea izolată sau continuă
- condiţii de zgomot şi robusteţea sistemului
- domeniul de aplicabilitate
- timpul de operare, care poate fi în timp real, cu întârziere sau off-line
- procentajul de recunoaştere
- costul.

Rezultate şi discuţii

Semnalul vocal reprezintă convoluţia în timp a semnalului excitaţie produs de vibraţia corzilor vocale şi răspunsul în timp al filtrului reprezentat de tractul vocal. Astfel, supus analizei vom avea un semnal convolvat. Separarea celor doua semnale, precum şi analiza parametrilor în domeniul temporal este imposibilă. Prin urmare, problema care se pune este trecerea semnalului în alt domeniu în care componentele se combină liniar, caz în care separarea ar fi simplă. Un astfel de domeniu este domeniul frecvenţa. Trecerea din domeniul temporal în domeniul frecvenţa se face cu ajutorul transformatei Fourier. În urma unei transformări liniare a unui semnal combinat liniar se obţine o combinaţie tot liniară. În consecinţă, dacă un semnal este obţinut prin combinaţia liniară a două semnale de frecvenţe diferite, în domeniul frecvenţa combinând liniar componentele, vom obţine semnalul ca o combinaţie liniară a componentelor spectrale. Pentru a demonstra că cele două componente sunt uşor de separat în domeniul frecvenţa, vom lua un exemplu în care un semnal de joasă frecvenţă $x(n)$ este perturbat de zgomotul de înaltă frecvenţă $w(n)$ • $y(n) = x(n) + w(n)$. Aplicând transformata Fourier în domeniul frecvenţa, vom avea câte o componentă separată la frecvenţa celor două

semnale. Prin eliminarea frecvenţelor nedorite şi transformarea inversă înapoi în domeniul timp se obţine semnalul curăţat de zgomot. În cazul semnalului vocal componentele nu sunt combinate liniar. Astfel, pentru a separa liniar componentele trebuie trecut şi semnalul vocal într-un domeniu, unde componentele se combină liniar. Fie *s(n)* semnalul vocal obţinut prin convoluţia excitaţiei corzilor vocale *e(n)* şi a răspunsului la impuls al funcţiei de transfer a tractului vocal $\Phi(n)$.

$$s(n) = e(n) \otimes \Phi(n) \tag{1}$$

Aplicând asupra acestei egalităţi transformata Fourier, convoluţia devine înmulţire şi, prin urmare, vom avea:

$$F\{s(n)\} = F\{e(n) \otimes \Phi(n)\} = F\{e(n)\} \cdot F\{\Phi(n)\} = E(\omega) \cdot \Phi(\omega) \tag{2}$$

Pentru o combinare liniară avem nevoie de adunare. Trecerea din multiplicativ în aditiv se face prin logaritmare şi, prin urmare, vom avea:

$$C(\omega) = \log\left[F\{s(n)\}\right] = \log\left[E(\omega) \cdot \Phi(\omega)\right] = \log\left[E(\omega)\right] + \log\left[\Phi(\omega)\right] \tag{3}$$

Am ajuns într-un domeniu în care componentele sunt combinate liniar. De acum încolo vom considera $C(n)$ ca semnal. Pentru a putea trece într-un domeniu în care componentele $\Phi(n)$ sunt uşor de separat, putem aplica iarăşi transformata Fourier, trecând în domeniu „frecvenţa":

$$c(n) = \frac{1}{2\pi} \cdot \int_{-\infty}^{+\infty} C(\omega) \cdot e^{-j\omega n} \cdot d\omega \tag{4}$$

Funcţia $C(\omega)$ fiind o funcţie reală pară, relaţia (4) este echivalentă cu:

$$c(n) = \frac{1}{2\pi} \cdot \int_{-\infty}^{+\infty} C(\omega) \cdot e^{j\omega n} \cdot d\omega, \tag{5}$$

deoarece

$$\int_{-\infty}^{+\infty} C(\omega) \cdot e^{-j\omega n} \cdot d\omega = \int_{-\infty}^{+\infty} C(\omega) \cdot \left(\cos(\omega n) - j \cdot \sin(\omega n)\right) d\omega = \int_{-\infty}^{+\infty} C(\omega) \cdot \cos(\omega n) \cdot d\omega$$

$$\int_{-\infty}^{+\infty} C(\omega) \cdot e^{j\omega n} \cdot d\omega = \int_{-\infty}^{+\infty} C(\omega) \cdot \left(\cos(\omega n) + j \cdot \sin(\omega n)\right) \cdot d\omega = \int_{-\infty}^{+\infty} C(\omega) \cdot \cos(\omega n) \cdot d\omega$$

unde

$$\int_{-\infty}^{+\infty} C(\omega) \cdot \sin(\omega n) \cdot d\omega = 0$$

Aplicând amândouă relaţiile, se obţine ca rezultat semnalul de start. În cazul în care din domeniul spectral se alege aplicarea transformării Fourier directe, domeniul în care se ajunge nu are semnificaţie fizică. Dacă se alege transformarea Fourier inversă, se ajunge într-un domeniu temporal. Acest domeniu temporal nu mai este identic cu domeniul temporal de unde am pornit, de aceea noţiunile utilizate pentru caracterizare se modifică (noţiuni de natură spectrală într-un domeniu temporal). Astfel, noul domeniu va avea denumirea de *domeniu cepstral*. Denumirile din noul domeniu se determină prin corespondenta cu domeniul spectral, însă inversând literele din prima silabă. Deci, se definesc următoarele denumiri: *spectrum* $\Leftrightarrow$ *cepstrum*, *frecvenţă* $\Leftrightarrow$ *cvefrenţă*, *armonica* $\Leftrightarrow$ *ramonica*, *filtrare* $\Leftrightarrow$ *liftrare*. În acest domeniu cepstral, componentele combinative ale semnalului original pot fi separate prin metode simple, cum ar fi liftrarea.

Schema-bloc a întregului proces este prezentată în figura următoare:

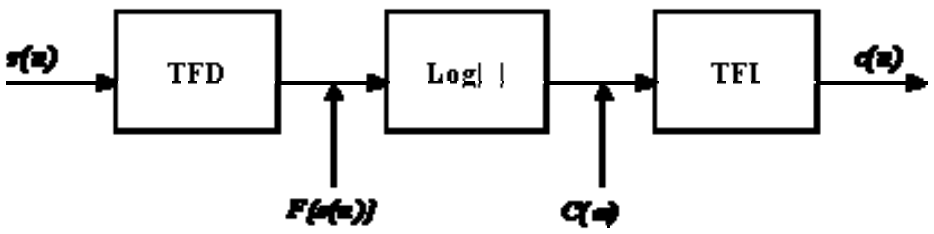

Fig.1. Schema-bloc de determinare a cepstrului.

Spectrul semnalului vocal este compus prin multiplicare dintr-o *anvelopă* de variație lentă şi *pulsuri*, o componentă cu variație rapidă. Spectrul logaritmic are forma asemănătoare, însă el reprezintă combinația liniară a celor două componente. În cepstrum (spectrul spectrului) se pot vedea separate cele două componente, una de joasă cvefrență, rezultată din componenta lent variabilă a anvelopei, şi componenta de cvefrență înaltă, rezultată din componenta rapid variabilă a spectrului. Distanța în cvefrența dintre două impulsuri, sau distanța de la origine la primul impuls, se poate interpreta ca şi perioada acestor impulsuri, deci se poate obține frecvența fundamentală de rezonanță a corzilor vocale. Această metodă este folosită pentru determinarea frecvenței fundamentale şi pentru determinarea caracterului vocalic sau nevocalic al cadrului de semnal analizat.

$$c_x(n) = c_e(n) + c_g(n) \tag{6}$$

Bineînțeles, această metodă poate fi aplicată pentru determinarea frecvenței fundamentale numai în cazul în care cele două componente combinatorice sunt destul de distanțate între ele. În cazul semnalelor de origine vocalică (vocale), metoda se justifică, însă în cazul sunetelor nevocalice (consoane), metoda este inadecvată, pentru că impulsurile, dacă există, se confundă cu zgomotul cepstral de înaltă cvefrență. Pe baza semnalului cepstral se pot obține, independent, spectrul de variație lentă sau cel de variație rapidă, prin operația de liftrare şi transformare în domeniul frecvența. De exemplu, dacă se dorește anvelopa spectrală, din relația (6) se elimină $c_g(n)$ şi se aplică transformata Fourier directă.

Concluzii

Odată determinați coeficienții cepstrali pentru cadrele dintr-un semnal vocal, se pune problema comparării acestora cu alt set de coeficienți ai cuvintelor etaloane din vocabularul analizat. Compararea se face calculând distanța între cele două cepstre. Pentru aceasta folosim relația lui Parseval, conform căreia energia semnalului periodic, calculată din eşantioane, este egală cu energia calculată din componentele spectrale înmulțite cu un factor. Considerând semnalul „temporal" cel spectral, se poate scrie egalitatea:

$$\sum_{n=-\infty}^{+\infty} c_n^2 = \frac{1}{2\pi} \cdot \int_{-\pi}^{\pi} \left[\log\left|H(\omega)\right|\right]^2 d\omega \tag{7}$$

În formula

$$s(n) = \sum_{p=1}^{p} a_p \cdot s(n-p) + G \cdot u(n) \tag{8}$$

demonstrată anterior se consideră factorul de amplificare *G=1,* iar suma coeficienților cepstrali făcându-se de la *n=1*, dat fiind faptul că ceilalți coeficienți sunt nuli. Distanța dintre cele două spectre se poate calcula ca fiind energia conținută în spectrul obținut prin scădere componentă cu componentă, care este identică cu o distanță euclidiană:

$$E_{sp} = \frac{1}{2\pi} \cdot \int_{-\pi}^{+\pi} \left|H_1(\omega) - H_2(\omega)\right|^2 d\omega \tag{9}$$

Prin logaritmare, egalitatea (9) devine:

$$E_{sp} = \frac{1}{2\pi} \cdot \int_{-\pi}^{\pi} \log\left|H_1(\omega) - H_2(\omega)\right|^2 \cdot d\omega \tag{10}$$

Din (9) se observa că energia E_{sp} este egală cu energia semnalului cepstral diferențial, care, fiind de natură temporală, se poate calcula şi ca energia semnalului diferență:

$$E_{sp} = \sum_{n=1}^{+\infty} (c_{1n} - c_{2n})^2 \tag{11}$$

Se poate scrie relația:

$$\sum_{n=1}^{+\infty} (c_{1n} - c_{2n})^2 = \frac{1}{2\pi} \cdot \int_{-\pi}^{\pi} \log\left|H_1(\omega) - H_2(\omega)\right|^2 \cdot d\omega \tag{12}$$

Prin urmare, distanţa euclidiană între două semnale cepstrale este o distanţă care poate exprima o diferenţă spectrală. În practică distanţa se calculează între coeficienţi până la ordinul *P*. Generalizând, distanţa cepstrală se defineşte:

$$d_{cep} = \sum_{n=1}^{p} (c_{1n} - c_{2n})^2 \qquad (13)$$

A fost studiat efectul ponderării în calculul distanţei cepstrale. Ponderarea îmbunătăţeşte performanţele calculului distanţei, ale recunoaşterii, deoarece coeficienţii cepstrali de ordin diferit au importanţă diferită din punctul de vedere al tractului vocal. Analiza cepstrală este o analiză specială, prin care efectul compus al excitaţiei şi al funcţiei de transfer a tractului vocal se separă cu o eficienţă mai bună, ca în cazul analizei spectrale sau liniar predictive.

Bibliografie:

1. Обжелян Н.К., Трунин–Донской В.Н. Речевое общение в системах „Человек – ЭВМ". - Кишинёв: Штиинца, 1985.
2. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. - Киев: Наукова Думка, 1987.
3. Rabiner L.R., Juang B.H. Fundamentals of speech recognition. - Prentice Hall, 1993.
4. Speech Analysis FAQ http://svr-www.eng.cam.ac.uk/~ajr/SA95/SpeechAnalysis.html
5. Microsoft Speech API Help. http:// research.microsoft.com.
6. Методы автоматического распознавания речи: В 2-х книгах. Пер. с англ. / Под ред. У.Ли. - Москва: Мир, 1983.

Prezentat la 31.01.2011