

# TESTAREA EXPERIMENTALĂ A UNOR ALGORITMI DE COMPRESIE A DATELOR

drd. ing. Cristian-Valentin Eremia

cristian\_valentin2003@yahoo.com

prof. dr. ing. Mihai Tertişco

mihai\_tertisco@yahoo.com

Universitatea Politehnica Bucureşti

**Rezumat:** Lucrarea prezintă rezultatele testării experimentale a 5 algoritmi de compresie a datelor (Shannon-Fano, LZW, Huffman Standard, Huffman Dinamic, Compresie Aritmetică). Testarea a avut scop evidenţierea performanţelor acestor algoritmi de compresie în funcţie de caracterul datelor (text, imagine, sunet, executabil, librării dinamice, C++, Pascal).

**Cuvinte cheie:** Shannon-Fano, LZW, Huffman Standard, Huffman Dinamic, Compresie Aritmetica

**Abstract:** The paper presents results of experimental testing of 5 data compression algorithms (Shannon-Fano, LZW, Standard Huffman, Dynamic Huffman, Arithmetic Compression). Testing was intended to highlight the performance of compression algorithms depending on the nature of the data (text, image, sound, executable, dynamic libraries, C++, Pascal).

**Keywords:** Shannon-Fano, LZW, Standard Huffman, Dynamic Huffman, Arithmetic Compression

## 1. Introducere

Scopul compresiei de date este acela de a îmbunătăţi performanţele operaţiilor ulterioare efectuate asupra acestor date: stocare, transmisie etc. Toate sistemele de compresie necesită doi algoritmi: unul pentru comprimarea datelor la sursă şi altul pentru decomprimarea datelor la destinatar. Prin compresie de date se urmăreşte ca, folosind anumite procedee, să se realizeze o trecere (transformare) a unui fişier  $F$ , de lungime  $L$  biţi, într-un fişier  $F_c$ , de lungime  $L_c < L$ . Deci decompresia transformă un fişier comprimat  $F_c$  într-unul identic cu cel original  $F$  sau într-unul apropiat de  $F$ , din punct de vedere al conţinutului. Algoritmii de compresie şi respectiv decompresie prezintă unele asimetrii privind cerinţele impuse performanţelor acestora în funcţie de domeniul de aplicaţie al datelor din fişiere. Spre exemplu, în multe aplicaţii, un document multimedia va fi codificat o singură dată (la sursă) şi decodificat de mii de ori (la destinaţie). Deci, această asimetrie oferă posibilitatea ca algoritmul de codificare să fie lent, în timp ce algoritmul pentru decodificare trebuie să fie rapid. În multe sisteme de compresie s-au făcut eforturi pentru ca decodificarea să se facă cât mai simplu şi să fie cât mai rapidă chiar şi cu preţul încetinirii şi complicării codificării. Acest lucru nu este valabil şi pentru aplicaţiile în timp real (spre exemplu la o video conferinţă). În timp real codificarea se face cu algoritmi diferiţi de cei folosiţi în cazul sistemelor de memorare (stocare) a unor fişiere cu date video, caz în care algoritmii pot fi mai lenţi decât în sistemul de timp real [1].

O altă asimetrie constă în faptul că procesul codificare/decodificare nu trebuie totdeauna să fie strict inversabil. Adică, atunci când se comprimă un fişier – program sau fişier text, dorim ca după decomprimare să-l obţinem exact pe cel original până la ultimul bit. În cazul fişierelor multimedia această cerinţă nu se pune. De obicei se acceptă să avem după decodificare un semnal puţin diferit de original. Atunci când ieşirea decodificată este identică cu intrarea originală sistemul este cu pierderi. Sistemele cu pierderi sunt importante deoarece acceptarea unui număr mic de informaţii pierdute poate oferi un avantaj imens în termenii de rată de comprimare. Indiferent dacă fişierele conţin texte, imagini, sau reprezentări ale sunetelor, la baza lor stă un alfabet.

Un alfabet se defineşte ca fiind totalitatea simbolurilor elementare folosite într-un fişier. Alfabetul se caracterizează prin lungime exprimată ca număr de simboluri.

Un cuvânt este o succesiune de simboluri elementare din alfabet grupate astfel încât să aibă o anumită semnificaţie. Cuvântul se caracterizează prin: număr de simboluri; delimitator de început şi delimitator de sfârşit; elementul corespondent dintr-o mulţime.

Un limbaj este alcătuit din mulțimea de cuvinte și regulile de folosire a acestora.

Compresia datelor este folosită și în cadrul sistemelor de stocare ori de transmisie la distanță, a datelor de tip text sau a datelor de tip sunet ori de tip imagine reprezentate în format digital. Explozia dimensională a fișierelor de date are loc în primul rând în cazul sistemelor multimedia care comportă transmisia și prelucrarea unui volum mare de informații în rețelele de calculatoare.

Toate sistemele de compresie conțin un codor și un decodor. Codorul realizează conversia datelor provenite de la sursă, în date comprimate, iar decodorul încearcă, la recepție, să reconstituie datele inițiale pe baza datelor comprimate. Datele reconstituite prin decodare fie coincid cu datele emise de sursă, fie diferă într-o măsură ne semnificativă pentru scopul propus. Această diferență între lungimea în biți a fișierului comprimat și a celui necomprimat trebuie evaluată și sistemul de compresie se ajustează funcție de scopul propus. Pentru a evalua calitățile unui sistem de compresie se definește o mărime numită raport de compresie:

$$\frac{\text{Dimensiunea datelor emise de sursă în biți}}{\text{Dimensiunea datelor comprimate în biți}} = r \quad (1.1)$$

Un raport de compresie "2 la 1", semnifică faptul că datele comprimate au jumătate din dimensiunea datelor emise de sursă. Un raport de compresie mare va desemna un sistem de compresie mai bun.

Procedurile de compresie de date se utilizează de mai multă vreme, fiind la început legate de stocarea pe un suport magnetic a textelor redactate în format electronic. Chiar și în acest domeniu evoluția a fost foarte spectaculoasă, performanțele algoritmului de compresie îmbunătățindu-se în ritm accelerat. Dar aceste tehnici care permit prin operația inversă de decompresie, refacerea exactă (fără pierderi) a ansamblului original oferă un grad redus de compresie, deoarece se bazează doar pe eliminarea redundanței naturale a sursei de informație, care chiar dacă există, are o valoare limitată. Rapoartele mult mai mari de compresie se pot obține dacă se renunță la refacerea exactă a semnalului informațional, dar atunci trebuie acceptat un compromis între factorul de compresie și precizia cu care se realizează refacerea.

Scopul acestei lucrări este să se realizeze:

- Studiul teoriei frecvențiale Shannon de compresie a fișierelor de tip TXT;
- Implementarea unui sistem ilustrativ bazat pe algoritmul Shannon–Fano, Huffman etc. de compresie;
- Testarea experimentală a principalelor metode de compresie a fișierelor de date și evaluarea comportamentului acestora în cazul diverselor tipuri de fișiere: **TXT**, **EXE**, **C++**, **Pascal**, **sunet**, **video**. Evaluarea metodelor se face pe baza ratei de compresie.

## 1.1. Teorema lui Shannon privind fișierul comprimat

Conceptele de entropie, cantitate de informații și redundanță au fost definite la începutul anilor '40, adică odată cu primele noțiuni de Teoria Informației. Ideea de bază care a declanșat numeroase cercetări în domeniul Compresiei Datelor a fost următoarea: *dacă s-ar cunoaște probabilitatea ce apare într-un set de date, atunci aceste simboluri ar putea fi recodificate în așa fel încât lungimea totală a noilor coduri să fie inferioară lungimii originale a setului de date*. În termeni de compresie, acest efect se numește *minimizarea redundanței setului de date*.

**Redundanța empirică** se evaluează prin diferența între numărul de biți inițial  $N_0$  și numărul de biți din fișierul comprimat  $N_c$ .

**ENTROPIA** unui simbol exprimă cantitatea de "nedeterminare" pe care o elimină "prezența" simbolului respectiv în fișier. Spre exemplu, dacă din alfabet este "șters" unul din simbolurile de frecvență maximă, practic nu se mai înțelege ce vrea să exprime conținutul informațional al fișierului. Deci, cantitatea de informație conținută în simbol, respectiv aportul informațional al simbolului la cantitatea de informație a întregului fișier, este mare, dacă

simbolul are frecvența mare și deci **entropia aceluiași simbol este mică**.

Relația de definiție, propusă de SHANNON pentru ENTROPIA  $H(sk)$  a unui simbol  $sk$  din alfabet este[2]:

$$H(sk)=-f(sk)\log_2f(sk), \quad (1.2)$$

în care  $\log_2$  este logaritmul în baza 2.

**De remarcat** că, întrucât frecvențele sunt aceleași atât pentru simbolurile din fișierul comprimat cât și necomprimat, entropiile simbolurilor sunt aceleași dar lungimile codurilor compresate sunt diferite în fișierul comprimat și necomprimat.

## 1.2. Teorema Shannon privind lungimea și entropia fișierului comprimat:

### Observație:

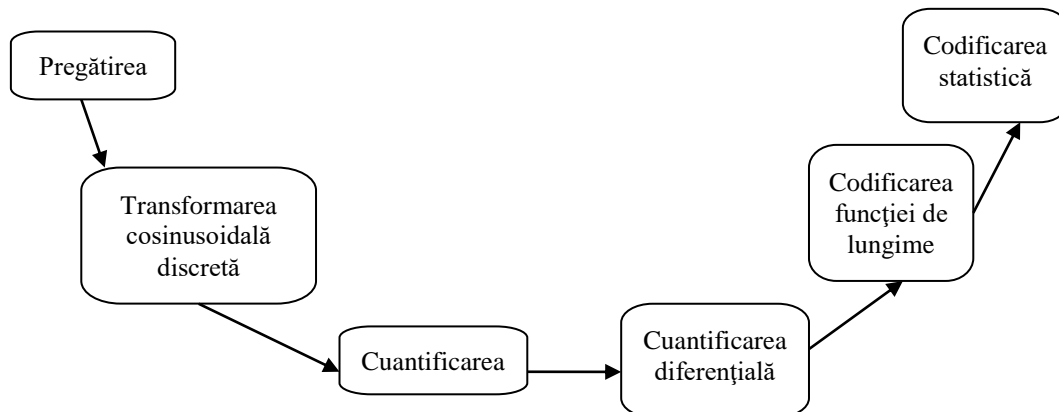
**Nici o metodă de compresie nu poate furniza o lungime  $L_c$ , a fișierului comprimat, mai mică decât  $H$  – entropia fișierului.**

Compresia de date este operația prin care se reprezintă compact datele furnizate de o anumită sursă. În funcție de sursa care le generează, forma și tipul de date variază:

- text,
- imagine,
- semnal vocal,
- semnal video.

Când ne referim la multimedia ne referim la combinarea dintre două sau mai multe media continue, adică media care trebuie să se desfășoare într-un interval bine definit, de obicei folosind interacțiunea cu utilizatorul. În practică cele două media sunt **audio** și **video**. O undă audio este o undă cu specific acustic.

## 1.3. Standardul JPEG pentru compresia imaginilor în tonuri continue



**Figura 1. Structura și conținutul procedurii, în 6 pași de codificare a imaginii RGB cu JPEG**

Standardul JPEG (Joint Photographic Experts Group) pentru comprimarea imaginilor în tonuri continue, a fost dezvoltat de experții în fotografie. Este important pentru multimedia deoarece la o primă aproximare, standardul multimedia pentru filme, MPEG este codificarea JPEG a fiecărui cadru separat, plus câteva caracteristici pentru comprimarea între cadre și detectarea mișcării. JPEG este definit în Standardul Internațional 10918. JPEG are patru moduri și multe opțiuni, fiind folosit în mod normal pentru codificarea imaginilor video de 24 biți RGB. Ne vom ocupa mai mult de modul secvențial cu pierderi. În figura 1.1 este prezentată structura și conținutul procedurii, în 6 pași de codificare, cu JPEG, a unei imagini **RGB**.

## 1.4. Standard MPEG (Motion Picture Experts Group)

Algoritmii principali din familia MPEG sunt destinați pentru compresia video și sunt standarde internaționale din 1993. Deoarece filmele conțin atât imagini cât și sunete, MPEG le poate comprima pe amândouă. Primul standard finalizat a fost MPEG-1. Scopul lui a fost de a produce ieșiri video de calitate video-recorder-elor (352x240) folosind o rată de biți de 1,2 Mbps. Video necomprimat poate ajunge la 472 Mbps, reducerea lui la 1,2 Mbps este însemnată, chiar și la această rezoluție scăzută. MPEG-1 poate fi transmis pe linii torsadate la distanțe modeste. Următorul standard din familia MPEG a fost MPEG-2 care a fost proiectat inițial pentru comprimarea video-ului de calitate de difuzare între 4 și 6 Mbps. Mai târziu MPEG a fost extins pentru a suporta rezoluții înalte. MPEG-4 este pentru video-conferințe de rezoluție medie cu cadre de viteză scăzută, (10 cadre/sec) și la lărgimi scăzute de banda (64 Kbps). Principiile de bază ale lui MPEG-1 și MPEG-2 sunt similare dar detaliile sunt diferite.

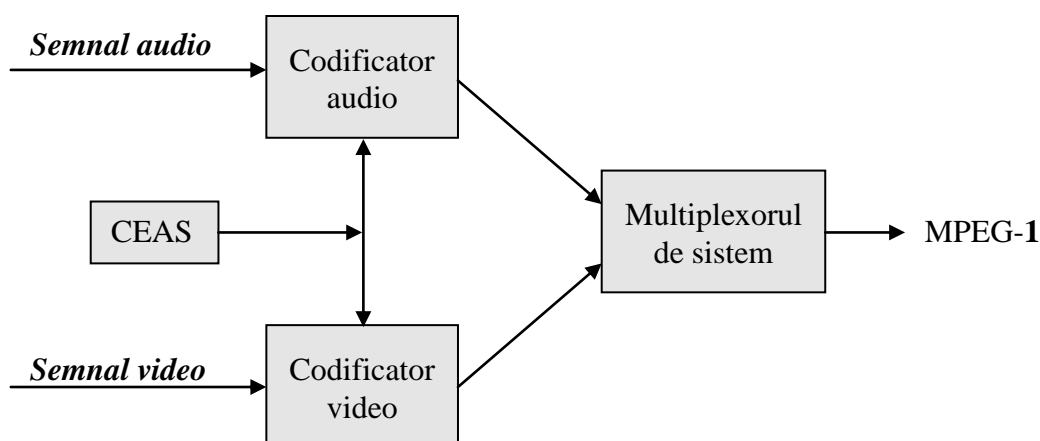


Figura 2. Structura standardului MPEG-1

Compresia audio MPEG este făcută prin eșantionarea formei de undă la 33 kHz, 44,1 kHz sau 48 kHz. Ea poate să gestioneze mono, stereo disjunct (fiecare canal separat) sau stereo reunit (este exploatată redundanța între canale). Este organizată pe trei nivele, fiecare dintre ele aplicând optimizări suplimentare pentru a obține o compresie mai mare. Nivelul 1 este schema de bază. Acest nivel este folosit, de exemplu, în sistemul de bandă DCC. Nivelul 2 adaugă schemei de bază alocarea avansată de biți. Este folosit pentru CD-ROM-uri audio și de piste sonore ale filmelor. Nivelul 3 adaugă filtre hibride, cuantificare neuniformă, codificare Huffman și alte tehnici avansate.

Audio MPEG poate comprima un CD de muzică până la 96Kbps fără o pierdere perceptibilă a calității. Numărul diferă de la o muzică la alta pentru că rata de semnal-zgomot diferă. Comprimarea audio este realizată prin transformata Fourier rapidă a semnalului audio pentru a-l converti din domeniul timpului în cel al frecvenței. Spectrul realizat este divizat în 32 benzi de frecvență fiecare procesată separat. Atunci când există două canale stereo, redundanța inerentă în a avea două surse audio suprapuse este de asemenea exploatată. Fluxul audio MPEG-1 rezultat este ajustabil de la 32Kbps la 448Kbps.

Obiectivul principal al cercetării în faza definitivării tezei de doctorat este testarea unor algoritmi de compresie pe diverse tipuri de fișiere de date cu scopul de a verifica experimental pentru care tipuri de fișiere se pretează, cel mai bine, diverșii algoritmi pentru compresii de date. Rezultatele experimentale ale acestei testări sunt conținute de acest capitol.

## 2. Testarea experimentală

În ceea ce privește programele de implementare a diverșilor algoritmi testați, le-am folosit pe cele care se găsesc prezentate în lucrarea recentă, intitulată „*Compresia datelor*”, publicată în anul 2003, de către domnul profesor universitar Dr. ing. Dan ȘTEFĂNOIU.

Există o diversitate foarte mare de algoritmi de compresie și variații ale acestora. Vom face o analiză comparată pe mai multe tipuri de fișiere, de dimensiuni diferite și cu diferiți algoritmi.

În analiză vom folosi următoarele 8 tipuri de fișiere: **dll**: biblioteci dinamice; **pas**: fișiere Pascal; **txt**: fișiere text; **exe**: fișiere executabile; **wav**: fișiere sunet; **bmp**: fișiere BitMap; **cpp**: fișiere C++; **img**: fișiere imagine.

Analiza o vom face folosind următorii algoritmi:

- A: compresia Huffman standard;
- B: compresia Shannon-Fano;
- C: compresie Huffman dinamic;
- D: compresie aritmetică de ordin n;
- E: compresie LZW

## 2.1. Calculul gradului de compresie

Relația de definiție a gradului de compresie  $g$ , utilizată pentru prelucrarea datelor experimentale este următoarea:

$$g = \left( 1 - \frac{L'}{L} \right) * 100 \quad (1.3)$$

Unde:

- $L'$  - lungimea fișierului compresat printr-un algoritm de compresie din mulțimea celor selectate pentru testare;
- $L$  - lungimea fișierului inițial necompresat.

## 2.2. Rezultatele testării celor 5 algoritmi pe 10 diverse tipuri de fișiere de lungimi diferite

Tabelul 1 - Rezultate pentru fișiere de tip dll (librării dinamice):

Fișier inițial (dim în bytes)	Huffman Standard	Shannon – Fano	Huffman Dinamic	Compresie Aritmetică	LZW
37440	33%	31%	34%	26%	49%
119056	26%	18%	19%	31%	52%
161552	33%	18%	19%	31%	48%
59904	32%	22%	23%	17%	54%
109424	30%	16%	17%	14%	47%
11776	31%	29%	34%	26%	39%
217088	43%	17%	18%	20%	50%
470528	27%	21%	16%	34%	51%
469504	33%	22%	16%	24%	52%
5632	33%	28%	38%	19%	49%
<b>Media</b>	33%	23%	24%	25%	50%

**Tabelul 2 - Rezultate pentru fișiere de tip PAS (surse Pascal)**

Fișier inițial (dim în bytes)	Huffman Standard	Shannon – Fano	Huffman Dinamic	Compresie Aritmetică	LZW
25840	38%	37%	38%	38%	55%
49249	43%	43%	44%	44%	59%
18845	41%	41%	42%	42%	61%
27712	35%	35%	36%	36%	57%
33336	36%	36%	37%	37%	55%
30238	37%	37%	38%	38%	58%
10901	35%	34%	36%	36%	53%
7494	33%	33%	35%	35%	50%
16738	36%	36%	38%	37%	55%
9432	33%	32%	34%	34%	51%
<b>Media</b>	37%	36%	38%	38%	56%

**Tabelul 3 - Rezultate pentru fișiere de tip TXT(fișiere text)**

Fișier inițial (dim în bytes)	Huffman Standard	Shannon – Fano	Huffman Dinamic	Compresie Aritmetică	LZW
19411	37%	36%	38%	38%	53%
36232	38%	37%	38%	39%	50%
42862	41%	41%	42%	42%	53%
4549	35%	35%	38%	39%	55%
36459	41%	41%	42%	42%	57%
20309	38%	38%	39%	42%	53%
23294	37%	37%	38%	38%	54%
10884	40%	40%	42%	42%	54%
60646	35%	37%	38%	38%	56%
90108	7%	34%	35%	35%	35%
<b>Media</b>	35%	37%	40%	41%	54%

**Tabelul 4 - Rezultate pentru fișiere de tip EXE (executabile)**

Fișier inițial (dim în bytes)	Huffman Standard	Shannon – Fano	Huffman Dinamic	Compresie Aritmetică	LZW
74192	37%	25%	29%	26%	22%
39280	30%	30%	31%	31%	29%
19456	29%	27%	32%	31%	29%
2708	9%	3%	19%	17%	13%
10774	9%	9%	14%	14%	11%
44032	25%	25%	26%	26%	22%
15656	17%	15%	20%	20%	16%
77312	49%	33%	36%	34%	33%
111376	36%	23%	25%	24%	19%
19729	17%	15%	19%	19%	16%
<b>Media</b>	26%	20%	25%	25%	22%

**Tabelul 5 - Rezultate pentru fișiere de tip WAV (sunet)**

Fișier inițial (dim în bytes)	Huffman Standard	Shannon – Fano	Huffman Dinamic	Compresie Aritmetică	LZW
12479	37%	15%	26%	26%	18%
55490	17%	6%	50%	45%	51%
1226	11%	18%	47%	30%	49%
7754	16%	16%	46%	50%	50%
4296	14%	12%	42%	42%	74%
12106	20%	13%	53%	22%	39%
25704	20%	20%	51%	51%	9%
22570	12%	12%	49%	14%	54%
33848	24%	2%	54%	51%	48%
28282	25%	12%	60%	47%	65%
<b>Media</b>	19%	14%	48%	38%	48%

**Tabelul 6 - Rezultate pentru fișiere de tip bmp (BitMap)**

Fișier inițial (dim în bytes)	Huffman Standard	Shannon – Fano	Huffman Dinamic	Compresie Aritmetică	LZW
2118	75%	75%	77%	79%	74%
582	56%	55%	61%	58%	57%
470	46%	46%	58%	52%	35%
2102	31%	30%	39%	35%	37%
578	60%	60%	67%	62%	67%
38462	20%	20%	22%	22%	22%
590	50%	50%	59%	55%	55%
339178	66%	54%	59%	54%	64%
578	53%	53%	59%	56%	33%
198	29%	29%	47%	34%	27%
<b>Media</b>	49%	47%	55%	51%	52%

**Tabelul 7 - Rezultate pentru fișierele de tip CPP (surse C++)**

Fișier inițial (dim în bytes)	Huffman Standard	Shannon – Fano	Huffman Dinamic	Compresie Aritmetică	LZW
15375	37%	37%	38%	38%	55%
40382	34%	34%	34%	35%	51%
23162	32%	32%	33%	33%	51%
478	14%	14%	34%	26%	25%
12501	32%	32%	33%	34%	47%
1354	25%	25%	36%	34%	35%
990	24%	24%	37%	34%	34%
35746	44%	43%	44%	45%	62%
405	16%	15%	40%	32%	21%
12234	35%	34%	36%	36%	59%
<b>Media</b>	30%	29%	37%	35%	44%

Tabelul 8 - Rezultate pentru fișiere de tip IMG (image)

Fișier inițial (dim în bytes)	Huffman Standard	Shannon – Fano	Huffman Dinamic	Compreseie Aritmetică	LZW
40410	26%	26%	27%	27%	27%
54696	29%	29%	31%	30%	65%
33822	34%	34%	35%	35%	38%
40028	17%	17%	19%	29%	51%
16900	26%	26%	29%	49%	48%
35640	18%	17%	19%	39%	36%
33622	17%	16%	18%	38%	41%
33230	19%	18%	20%	40%	44%
20894	19%	19%	21%	32%	70%
31652	18%	18%	20%	40%	39%
<b>Media</b>	23%	22%	25%	33%	46%

### 2.3. Rezultate ale analizei comparative a testului

În cadrul etapei de analiză comparativă a rezultatelor experimentale ale testului s-au realizat reprezentari grafice ,prin bargrafuri, pentru toate rezultatele comparației ratelor medii calculate de compresie. Aceste reprezentari grafice ale rezultatelor analizei comparative si concluziile desprinse din acestea sunt prezentate in continuare.

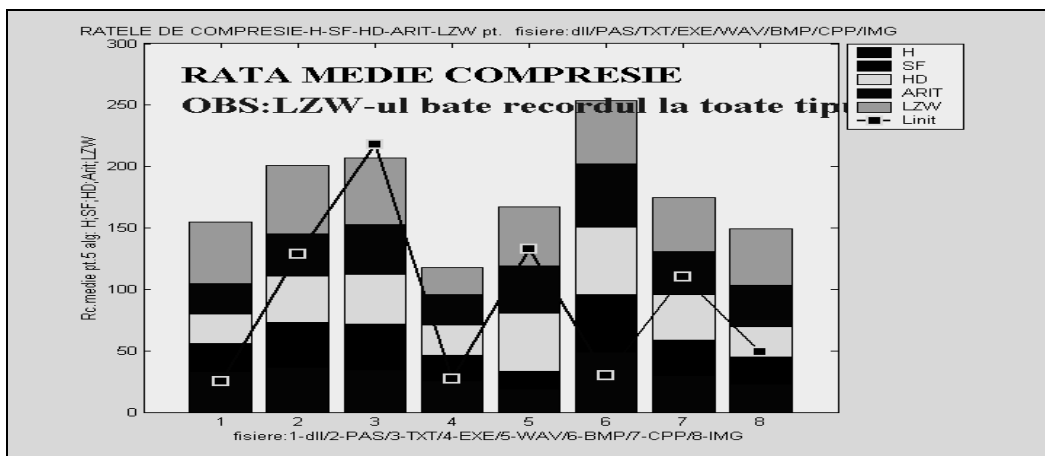


Figura 3. Valorile medii ale ratelor de compresie a 5 algoritmi și 8 tipuri de fișiere

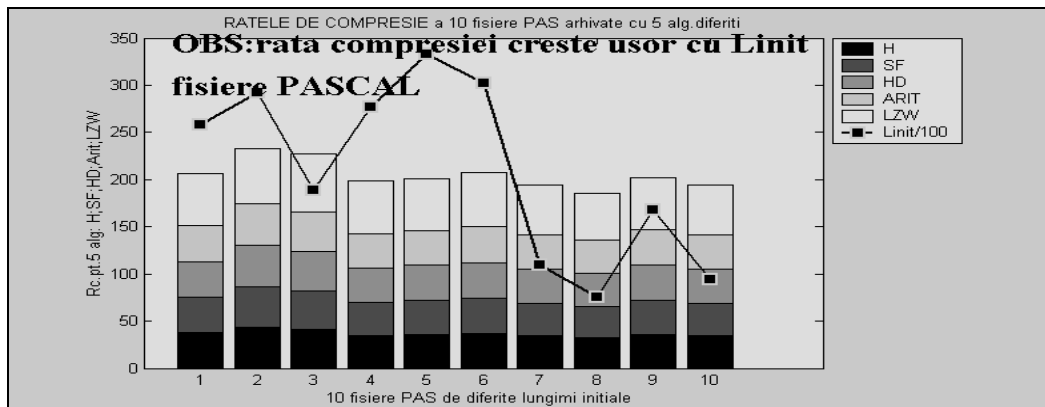


Figura 4. Cei 5 algoritmi testați pe 10 fișiere de același tip PASCAL dar de lungimi diferite prezentate în tablele de mai sus.



În figura 3, cele 8 fișiere sunt ordonate crescător ca lungime inițială. Concluzia care se desprinde este că, cea mai bună comportare o are algoritmul LZW care practic prezintă cea mai mare rată de compresie pentru toate fișierele, indiferent de lungimea acestora. În figurile 4 și 5 sunt prezentate rezultatele testării celor 5 algoritmi pe 10 fișiere de lungimi diferite dar de același tip, PASCAL respectiv WAV. Din figura 5 rezultă că la toți algoritmi, rata de compresie crește cu lungimea fișierelor și mai rezultă că algoritmi LZW și HD furnizează cele mai bune rezultate la compresia fișierelor de sunet.

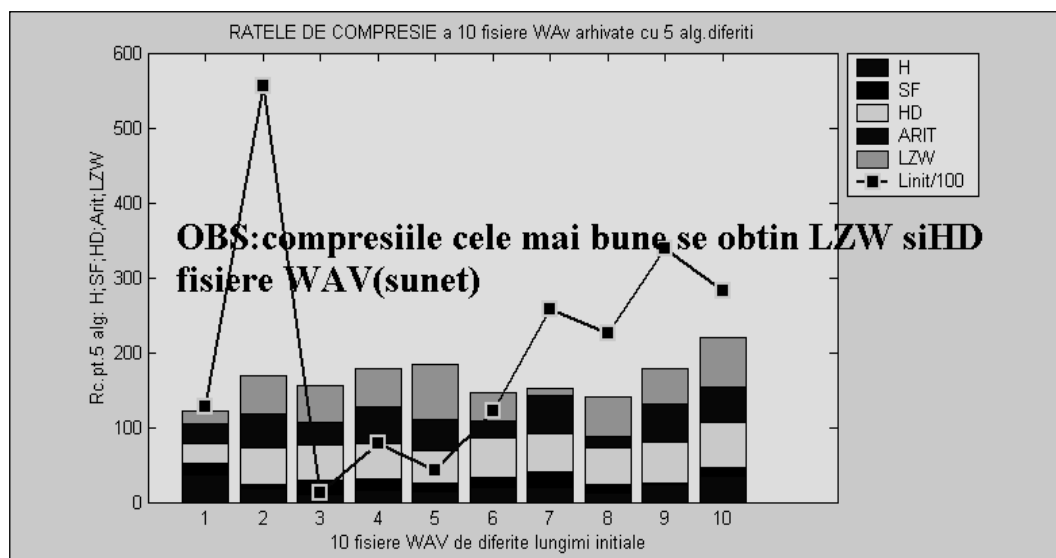


Figura 5. Cei 5 algoritmi testați pe 10 fișiere de același tip WAV dar de lungimi diferite

### 3. Concluzii

- Compresia datelor este folosită atât în cadrul sistemelor de stocare ori de transmisie la distanță, a datelor de tip text sau a datelor de tip sunet ori de tip imagine reprezentate în format digital. Aplicațiile dezvoltate în domeniul conducerii centralizate a unor roboți cu vedere artificială, precum și în cazul avioanelor de spionaj fără pilot este necesară transmisia datelor video într-o formă compresată pentru scurtarea duratei achiziției și transmisiei datelor la sol. Aceasta explică importanța acordată compresiei datelor de către cercetători.
- Tehnicile pentru compresia imaginilor sunt dominate de prelucrările de semnale în domeniul tehnologiei informației. Creșterea vertiginoasă a aplicațiilor care folosesc reprezentări 2D și 3D și în special transferul de fișiere grafice pe Internet au impus dezvoltarea de tehnici speciale și mai ales de standarde specializate. Procedurile de compresie de imagini sunt complexe, asociind mai mulți algoritmi. De aceea, rata distorsiunii este decisivă în alegerea nivelului de compresie, pentru că în funcție de aplicație se poate merge până la rate foarte înalte de compresie, dacă detaliile imaginii nu sunt importante sau se lucrează cu imagini binarizate. Un alt aspect important este acela a timpilor de rulare a algoritmilor de compresie/decompresie, anumite proceduri nefiind indicate pentru transmitere la distanță, ci doar pentru stocare, datorită timpului prea lung în care se face compresia. În fine, un aspect care nu poate fi neglijat este acela că există aplicații în care compresia de imagini trebuie asociată cu cea a semnalului audio și eventual a unor dale de proces sau fișiere text, ceea ce ridică probleme de împachetare a cadrelor și mai ales de armonizare a debitelor informaționale.
- Fișierele de sunet având dimensiuni relativ mici, comprimarea lor se realizează în bune condiții și în timp scurt folosind algoritmul Huffman Standard chiar dacă rata de compresie pentru aceste fișiere este destul de scăzută 35%.

## **BIBLIOGRAFIE**

1. **DIATCU, E., A. TERTIȘCO, F. IACOB, M. TACHE, Z. RACoviȚĂ:** Elemente fundamentale ale teoriei sistemelor și calculatoarelor, editura HYPERION XXI, 1997;
2. **STEFĂNOIU, D.:** Compresia datelor, editura Printech, 2003;