УДК 519.8: 004.934

А.М. НИЩУК
Науково-дослідний інститут воєнної розвідки, Київ, Україна, e-mail: *svs14@ukr.net*.

С.М. НІКОЛАЄВ
Науково-дослідний інститут воєнної розвідки, Київ, Україна, e-mail: *divan24@i.ua*.

О.М. РОМАНОВ
Науково-дослідний інститут воєнної розвідки, Київ, Україна, e-mail: *rolex@i.ua.*

МЕТОДИКА АНАЛІЗУ БІТОВИХ ПОТОКІВ НА ОСНОВІ ВИКОРИСТАННЯ ВІДСТАНІ ДАМЕРАУ–ЛЕВЕНШТЕЙНА ТА ІНШИХ МЕТРИК

Анотація. Проведено аналіз особливостей різних відстаней (метрик) під час розв'язання задач порівняння бітових потоків. На прикладі розглянуто застосування запропонованої методики для розв'язання задачі ідентифікації бітових потоків цифрового кодування мовлення. Показано, що у цьому випадку найбільш ефективною є метрика Дамерау–Левенштейна.

Ключові слова: бітовий потік, відстань, метрика, цифрове кодування мовлення.

ВСТУП. АКТУАЛЬНІСТЬ ДОСЛІДЖЕНЬ ТА АНАЛІЗ НАЯВНИХ РЕЗУЛЬТАТІВ

Післядетекторне оброблення сигналів зазвичай означає роботу з цифровими даними, представленими у вигляді бітових потоків (bitstreams). Це стосується більшості радіотехнічних та телекомунікаційних систем, які на фізичному рівні обробляють радіосигнали, а на канальному та інших рівнях — бітові потоки. Дослідники вживають такі синоніми терміна «бітовий потік»: цифровий потік; двійкова, дискретна або бітова послідовність; цифрове представлення сигналу; файл тощо. Автори пропонують послуговуватися саме терміном «бітовий потік», який має чіткий відповідник англійською мовою. Завдяки цьому можна коректно розрізнити цифрове оброблення сигналів (signal processing) та оброблення бітових потоків (bitstream processing).

Очевидно, що під час аналізу бітових потоків одним з важливих процесів є пошук. Цей пошук є основою відповідних систем в інтернеті, його також використовують у криптографії для виявлення періодичностей у бітових потоках. На практиці задачу порівняння та визначення статистичних властивостей бітових потоків розв'язують у різних технічних галузях: від радіотехніки і передавання інформації до біомедицини, оброблення зображень тощо. Основою порівняння бітових потоків є пошук періодичностей і повторень окремих фрагментів з використанням різних методів. Наприклад, у криптографії оцінювання бітового потоку здійснюють за допомогою тестів NIST [1]. Їх рекомендовано використовувати, щоб з'ясувати, чи можна вважати бітовий потік випадковим. Слід зазначити, що не завжди на основі цих тестів можна відрізнити архівовані файли від шифрованих [2]. До методики на основі зазначених тестів відносять перевірки бітового потоку на статистичні властивості, виявлення факту оброблення регістрами зсуву тощо. Однак, для порівняння двох та більше бітових потоків тести NIST застосовувати незручно і в багатьох випадках вони не забезпечують потрібного результату. Загалом для розроблення деякої системи ідентифікації або класифікації, яка ґрунтується на порівнянні бітового потоку з еталоном, потрібно аналізувати ці потоки на основі математичних виразів для обчислення певних відстаней (метрик). Аналіз особливостей зазначених метрик і є метою цієї статті.

Крім розглянутого підходу, що ґрунтується на тестах NIST, заслуговують на увагу методи, які дають змогу на основі аналізу бітового потоку визначати його період за допомогою автокореляційних та взаємокореляційних функцій [3], наявність і параметри завадостійких кодів [4, 5], поліноми скремблерів [6], наявність й параметри перемеження [7], тип вокодера [8]. Однак у більш загальному випадку ці методи не є універсальними та їхнє застосування обмежене. У [9] для пошуку періодичностей запропоновано використовувати функцію потужності періодичної складової, що має локальні максимуми навіть за наявності періоду бітового потоку, і яку не можна застосувати для порівняння бітових потоків у системах ідентифікації.

СУТЬ МЕТОДИКИ АНАЛІЗУ БІТОВИХ ПОТОКІВ

Методика аналізу бітових потоків може ґрунтуватися на їхньому порівнянні з еталонами шляхом обчислення певних математичних виразів. Автори пропонують проаналізувати суть та дослідити ефективність дев'яти метрик, наведених на рис. 1. Під метрикою розуміють показник ефективності процесу, що характеризується властивостями відстані [10]. Деякі метрики можна визначити більш простим терміном «відстань». Відстань — це міра відмінності двох послідовностей [10].

Проте у більш складних випадках для обчислення застосовують певний алгоритм і термін «відстань» не повною мірою характеризує суть обчислень. Тому далі для зручності вжито термін «метрика».

Відстань Хеммінга — це кількість позицій, в яких відповідні бітові значення для двох послідовностей однакової довжини є різними [11]:

$$D_H = \sum_{k=1}^{m} s_1[k] \oplus s_2[k],$$

де s_1 та s_2 — два бітові потоки, k — лічильник, m — довжина бітових потоків. Чим більша відстань Хеммінга, тим менша подібність бітових послідовностей між собою. Максимально можлива відстань max (D_H) становить m. На практиці алгоритм для обчислення відстані Хеммінга легко реалізувати. Його широко застосовують у теорії кодування та телекомунікаційних системах. Однак, ця метрика має два суттєві недоліки:

— бітові потоки повинні мати однакову довжину;

— за наявності зсувів у бітових потоках результати будуть неточними.

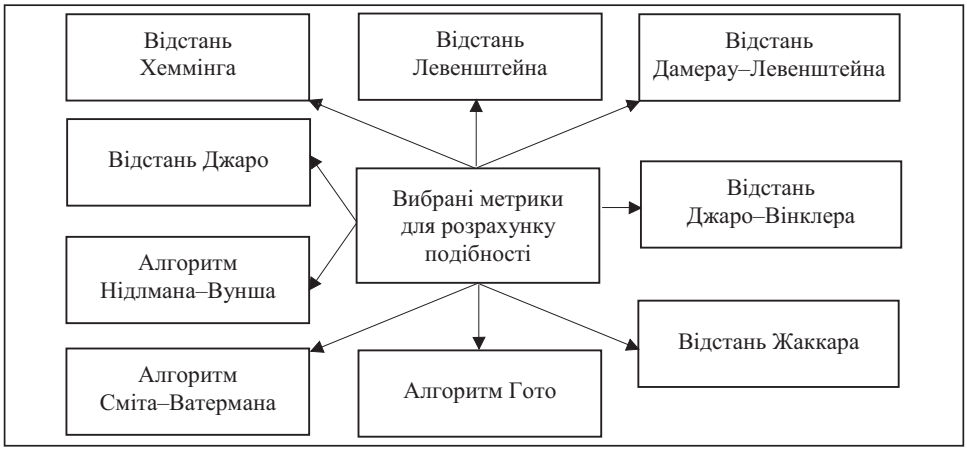

Рис. 1. Схематичне зображення досліджуваних метрик

Відстань Левенштейна — це метрика, що визначається мінімальною кількістю операцій вставки одного біта, видалення одного біта та заміни одного біта на інший, необхідних для перетворення однієї послідовності в іншу [12]. Її обчислюють за такою рекурентною формулою:

$$D_L = \min \begin{cases} D_L + 1^K, \\ D_L + 1^I, \\ D_L + \begin{cases} 0^E, & \text{if } s_1[m] = s_2[n], \\ 1^R, & \text{else}, \end{cases} \end{cases}$$

де s_1 та s_2 — два бітові потоки; m та n — довжини бітових потоків.

Класифікація дозволених операцій та штрафи на виконання операції:

R — заміна біта (штраф = 1);

I — вставка біта (штраф = 1);

K — видалення біта (штраф = 1);

E — збіг біта (штраф = 0).

Розрахунок метрики Левенштейна можна змінити шляхом корегування значення штрафів. Інакше кажучи, для кожної операції над бітами (заміна R , вставка I, видалення K) можна встановити відповідний штраф. На відміну від метрики Хеммінга метрика Левенштейна дає змогу порівнювати бітові потоки різної довжини, проте алгоритм її обчислення є більш складним у практичній реалізації. Для розрахунку цієї метрики можна скористатися методом динамічного програмування Вагнера–Фішера [13].

Відстань Дамерау–Левенштейна [14, 15] відрізняється від метрики Левенштейна тим, що до переліку дозволених операцій додано транспозицію T (два сусідні бітові символи міняються місцями):

$$D_{DL} = \min \begin{cases} D_{DL} + 1^K, \\ D_{DL} + 1^I, \\ D_{DL} + \begin{cases} 0^E, & \text{if } s_1[m] = s_2[n], \\ 1^R, & \text{else}, \end{cases} \\ D_{DL} + \begin{cases} 1^T, & \text{if } s_1[m-1] = s_2[n], \\ \infty, & \text{else}. \end{cases} \end{cases}$$

Метрика Джаро — це мінімальна кількість бітових перетворень, необхідних для заміни одного бітового потоку іншим [16]:

$$D_J = \begin{cases} 0, & c = 0, \\ \dfrac{1}{3}\left(\dfrac{c}{m} + \dfrac{c}{n} + \dfrac{c-t}{c} \right), \end{cases}$$

де m та n — довжини бітових потоків; c — кількість бітів, що збігаються; t — кількість транспозицій.

Метрику Джаро можна застосовувати до бітових потоків різної довжини. Вона забезпечує нормований результат від 0 до 1. Що меншим є значення результату, то більшим є збіг бітових потоків між собою.

Метрика Джаро–Вінклера відрізняється від метрики Джаро використанням коефіцієнта масштабування p [17]. Її визначають за такою формулою:

$$D_{JW} = D_j + l \cdot p \cdot (1 - D_j),$$

де D_j — значення відстані Джаро; l — значення кількості збігів початку бітового потоку (до чотирьох біт); p — постійний коефіцієнт масштабування (від 0.1 до 0.25).

Визначення подібності на основі метрики Джаро–Вінклера дає змогу більш точно обчислити подібність між бітовими потоками у разі наявності початкової синхрокомбінації, характерної для телекомунікаційних систем.

Алгоритм Нідлмана–Вунша [18] містить три етапи, а саме побудову ініціювальної матриці, заповнення матриці та пошук максимального вирівнювання. Заповнення здійснюють за такою формулою:

$$s_{ij} = d_{ij} + \max\left(s_{i-1,j-1}; \max_{k<j-1}(s_{i-1,k} - G); \max_{k<i-1}(s_{k,j-1} - G) \right),$$

де s_{ij} — елемент матриці; i — номер рядка; j – номер стовпця; d_{ij} — результат подібності між елементами; G — штраф (від 0 до 1).

Після цього здійснюють прохід по максимальним елементам матриці у зворотному напрямку. Отриманий маршрут відповідає оптимальному вирівнюванню D_{NV}.

Цей алгоритм слугує для глобального вирівнювання і є складнішим у реалізації, ніж попередні метрики. На практиці його здебільшого застосовують для аналізу та порівняння біологічних послідовностей.

Алгоритм Сміта–Ватермана [19], на відміну від алгоритму Нідлмана–Вунша, порівнює відрізки всіх можливих довжин та оптимізує міру подібності за всіма відрізками та всіма вирівнюваннями цих відрізків. Заповнення матриці здійснюють за таким виразом:

$$s_{ij} = \max(s_{i-1,j} + G;\ s_{i,j-1} + G;\ s_{i-1,j} + w;\ 0)\ \ (1 \le i \le n,\ \ 1 \le j \le m),$$

де w — оцінка збігу послідовностей; G — штраф (від 0 до 1).

Цей алгоритм, на відміну від попереднього, забезпечує локальне вирівнювання, що є ефективним способом пошуку подібності під час порівняння суттєво видозмінених бітових потоків.

Алгоритм Гото відрізняється від метрики Нідлмана–Вунша застосуванням штрафів за афінний розрив [20]. У цьому алгоритмі вважають, що одиночна подія вставки чи видалення великої кількості елементів є більш імовірною, ніж множина невеликих вставок чи видалень.

Метрика Жаккара [21] вимірює відмінність множин. Для двох бітових потоків її визначають за таким виразом:

$$D_Z = \frac{|\, s_1 \cap s_2 \,|}{\sum_{i=1}^{m} s_1[i] + \sum_{i=1}^{n} s_2[i] - |\, s_1 \cap s_2 \,|}.$$

Основними недоліками методу розрахунку метрики Жаккара, який ґрунтується на множині спільних бітів, є врахування лише статистики появи значень 0 та 1 і неможливість врахування позицій бітів у потоці.

Розглянемо ефективність застосування запропонованої методики на прикладі задачі ідентифікації бітових потоків шляхом порівняння з еталонами цифрового кодування мови. Для цього обмежимося дев'ятьма найпоширенішими типами цифрового кодування мови, основні параметри яких наведено у табл. 1.

Таблиця 1. Параметри цифрового кодування мовлення з низькошвидкісними алгоритмами кодування

Тип кодування	Швидкість кодування, кбіт/с	Довжина фрейму T, біт	Алгоритм кодування
Lyra	3	120	WaveRNN,MELP
LPC10	2.4	54	LPC
G.723.1	6.4	192	MP-MLQ
Codec2	2.4	48	LPC
MELP	2.4	54	MELP
Speex	3.95	80	CELP
Speex	5.95	120	CELP
G.729	8	80	ACELP

З огляду на те, що структури бітових потоків у паузах і за наявності мовлення суттєво різняться, для кожного типу цифрового кодування мовлення сформовано по два еталони — для паузи та мовлення відповідно. Для статистичного аналізу еталонів використано нормовану міру E, в якій для кожного біта періоду T підраховано кількість значень бітів, що відрізняються між собою. Функцію нормованої міри $E(i)$ визначають за формулою

$$E(i) = \sum_{j=2}^{K} x_{i1} \oplus x_{ij},$$

де $i = 1 \dots T$ — позиція біта в межах періоду; $K = 1000$ — кількість фреймів з періодом T.

Чим більше бітових значень між фреймами відрізняються, тим більшим є значення нормованої міри E. Графічні зображення нормованої міри E еталонів бітових потоків цифрового кодування мовлення наведено на рис. 2. На горизонтальній осі позначено позицію біта у межах періодіу T, на вертикальній — значення нормованої міри E.

З наведених зображень видно, що для цифрового мовленнєвого повідомлення, яке містить тільки паузу, довжина унікальних бітових синхропослідовностей є більшою ніж довжина мовленнєвих повідомлень. Для деяких типів цифрового кодування мовлення (MELP, LPC10, G.729) у разі передавання мовленнєвого сигналу значення нормованої міри візуально свідчать про відсутність повторень.

Далі той самий сеанс мовлення з паузами було закодовано дев'ятьма типами цифрового кодування мовлення і проведено порівняння відповідних бітових потоків з еталонами шляхом обчислення дев'яти метрик. Довжина цих бітових потоків відповідає довжині еталонів, сформованих раніше ($K = 1000$). Розраховані значення подібності за допомогою метрик сформовано відповідно до таких виразів:

$$D_{xy} = \min(D_H; D_L; D_{DL}),$$

$$D_{xy} = \max(D_J; D_{JW}; D_{SW}; D_{NW}; D_G; D_J),$$

де x — лічильник вхідних бітових потоків загальною кількістю n $(1 \le x \le n)$; y — лічильник еталонних бітових потоків $(1 \le y \le 18)$.

На основі розрахованих значень для кожної з дев'яти метрик побудовано матриці. У табл. 2 наведено матрицю подібності для метрики Дамерау–Левенштейна.

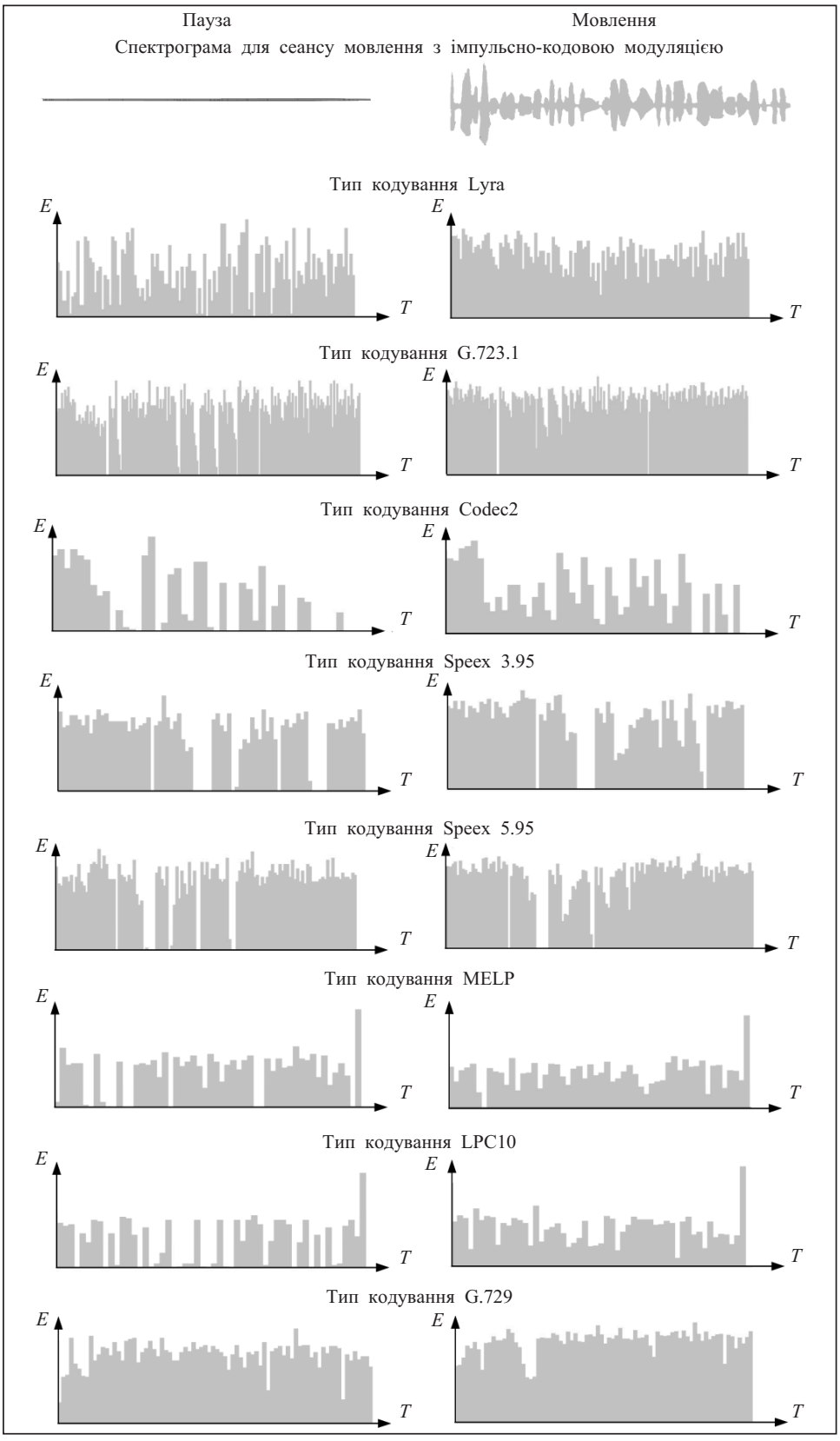

Рис. 2. Графічні зображення нормованої міри еталонів бітових потоків цифрового кодування мовлення

Таблиця 2. Матриця подібності для метрики Дамерау–Левенштейна

	0.86	0.80	0.84	0.85	0.97	0.90	0.79	0.82	0.88
0.85		Lyra	Lpc10	G.723.1	Codec2	Melp	Speex 3.95	Speex 5.95	G.729
0.81	Lyra	192	3	9	1	6	11	9	7
0.91	Lpc10	0	182	5	0	1	5	5	2
0.86	G.723.1	4	7	197	2	5	3	8	3
0.93	Codec2	0	2	2	196	0	8	0	3
0.92	Melp	5	1	7	1	191	0	2	1
0.92	Speex 3.95	3	1	0	0	2	190	8	2
0.82	Speex 5.95	12	12	4	0	2	7	170	0
0.67	G.729	25	8	8	2	5	15	6	137

На головних діагоналях матриць розміщено значення, що відповідають правильній ідентифікації. Для всіх проаналізованих метрик більшість бітових потоків визначені правильно, а діагональні елементи матриці містять максимальні значення. Серед усіх метрик для цього прикладу найбільш чітку ідентифікацію забезпечує метрика Дамерау–Левенштейна. На рис. 3 графічно зображено результат застосування методики для трьох з дев'яти метрик, що розраховані під час розв'язання задачі ідентифікації цифрового кодування мовлення.

Очевидно, що у разі розв'язання задачі аналізу (порівняння) бітових потоків на іншому рівні моделі OSI більш ефективною може бути інша метрика (у якої головна діагональ матриці, що відображена на графіку, містить максимальні значення (192, 182, 197, 196, 191, 190, 170, 137)).

Для оцінювання ефективності використання метрик застосовано показники точності та повноти [22]. Точність визначення типів цифрового кодування мовлення — це частка бітових потоків, що дійсно належать заданому типу цифрового кодування мовлення відносно всіх типів, віднесених системою до заданого типу. Повнота — це частка знайдених типів цифрового кодування мовлення, що належать заданому типу від загальної кількості типів цього кодування у тестовій вибірці. Ці значення розраховано на основі таблиці спряженості [23], яку складають для кожного типу (табл. 3).

Таблиця 3 містить оцінки, які дають змогу зрозуміти, скільки разів було прийнято правильне і скільки разів — неправильне рішення щодо бітових потоків заданого типу, а саме: TP — істино-позитивне рішення; TN — істино-негативне рішення; FP — хибно-позитивне рішення; FN — хибно-негативне рішення.

Тоді точність P і повноту R визначають у такий спосіб [23]:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}.$$

У табл. 2 точність відповідає відношенню діагонального елемента та суми всього рядка (наприклад, точність P для визначення бітового потоку G.729 є невисокою і становить 0.67). Повнота відповідає відношенню діагонального елемента та суми всього стовпця (наприклад, повнота R для визначення бітового потоку Codec2 є максимальною і становить 0.97).

Результівні значення точності і повноти, наведені в матриці та позначені жирним шрифтом (див. табл. 2), розраховано з використанням методу макроусереднення [22]:

$$P = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{c,i}} = 0.85, \quad R = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{i,c}} = 0.86.$$

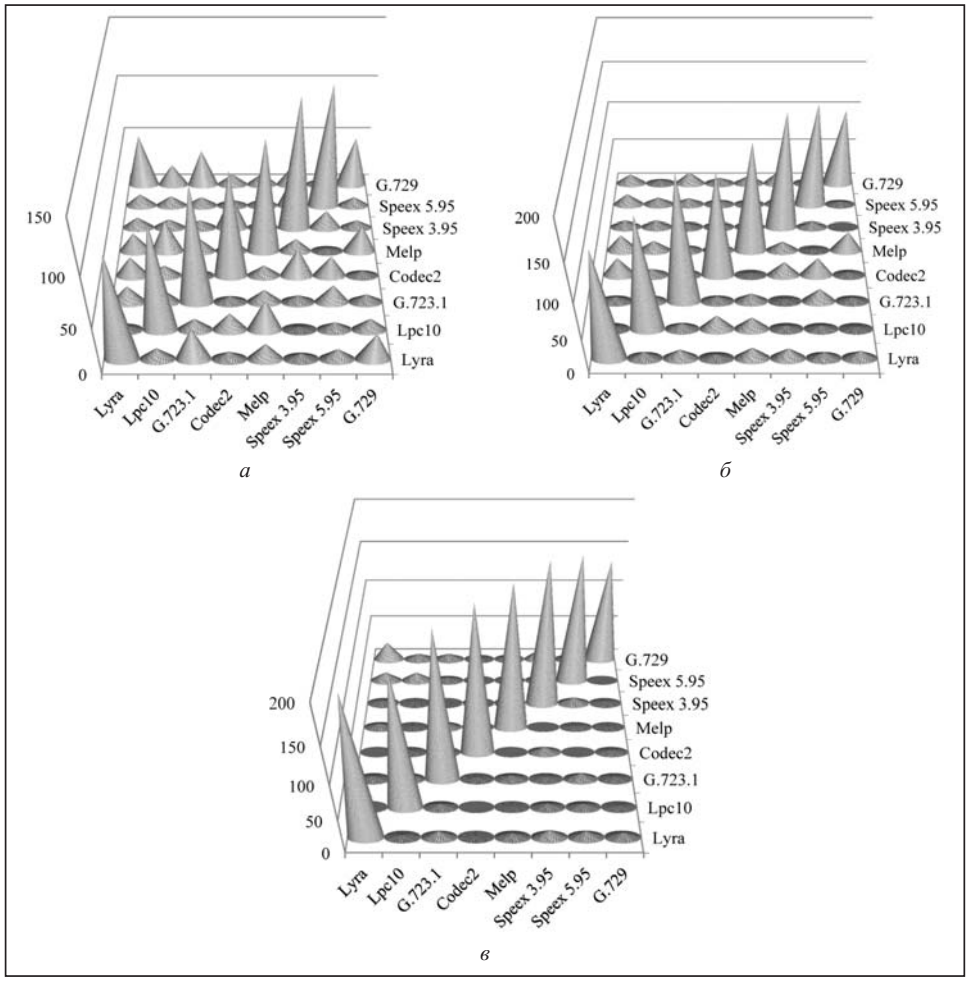

Рис. 3. Графіки розрахованих значень подібності на основі метрик: Жаккара (*а*), Хеммінга (*б*), Дамерау–Левенштейна (*в*)

Таблиця 3. Таблиця спряженості

Тип цифрового кодування мовлення		Експертна оцінка	
		Позитивна	**Негативна**
Оцінка системи	Позитивна	*TP*	*FP*
	Негативна	*FN*	*TN*

Для об'єднання інформації про точність і повноту застосовано збалансовану F-міру, яка є гармонійним середнім між точністю й повнотою та надає однакову вагу точності та повноті; для метрики Дамерау–Левенштейна вона становить $F = 2\dfrac{P \cdot R}{P + R} = 0.855$. Розраховані значення F-міри для всіх дев'яти метрик наведено на рис. 4 у вигляді гістограми.

Отже, методика аналізу бітових потоків на основі метрик для ідентифікації цифрового кодування мовлення з використанням метрики Дамерау–Левенштейна є більш ефективною, ніж у разі застосування інших метрик.

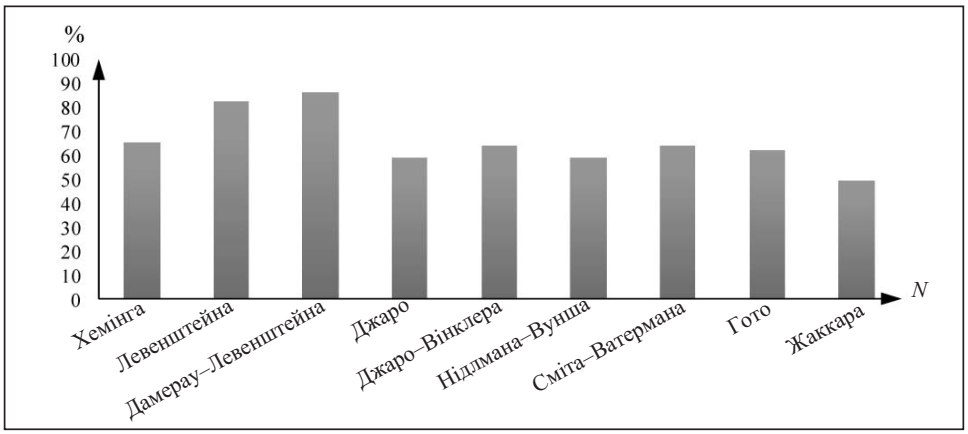

Рис. 4. Гістограма розрахованих у відсотках значень F-міри з використанням різних метрик

ВИСНОВКИ

Крім ідентифікації бітових потоків цифрового кодування мовлення розглянуті у статті метрики можна застосовувати в інших системах як основний підхід до аналізу, ідентифікації та класифікації бітових потоків. Наприклад, їх можна використовувати для порівняння протоколів мережевого або транспортного рівнів моделі OSI, визначення ефективності стиснення відео, порівняння псевдовипадкових послідовностей, під час аналізу стеганографічних повідомлень тощо. Водночас не обов'язково, що в таких системах використання метрики Дамерау–Левенштейна буде більш раціональним порівняно з іншими метриками.

Автори планують провести окремі дослідження та розрахунки для того, щоб для бітових потоків на кожному рівні моделі OSI визначити найбільш ефективну з дев'яти метрик, на яких ґрунтується запропонована методика. Відповідні результати підтвердять ефективність запропонованої методики аналізу бітових потоків для широкого класу практичних задач, пов'язаних з ідентифікацією та класифікацією.

З іншого боку, надалі потрібно більш детально дослідити ефективність застосування саме метрики Дамерау–Левенштейна на бітових потоках зі структурою, близькою до випадкової (неперіодичної).

СПИСОК ЛІТЕРАТУРИ

1. Rukhin A., Soto J., Nechvatal J., Smid M., Barker E., Leigh S., Levenson M., Vangel M., Banks D., Heckert A., Dray J., Vo S. A statistical test suite for random and pseudorandom number generators for cryptographic applications. National Institute of Standards and Technology, Gaithersburg, 2010. 131 p.
2. Козачок А.В., Спирин А.А., Голембиовская О.М. Алгоритм классификации псевдослучайных последовательностей на основе построения случайного леса. *Доклады Томского УСУРа*. 2020. Т. 23, № 3. С. 55–60. https://doi.org/10.21293/1818-0442-2020-23-3-55-60.
3. Романов О.М., Котюбін В.Ю. Алгоритми пошуку періодичностей у цифрових послідовностях з блоковим кодуванням за їх кореляційними властивостями. *Радіоелектроніка, інформатика, управління.* 2021. № 2 (57). С. 7–18. https://doi.org/10.15588/1607-3274-2021-2-1.
4. Marazin M., Gautier R., Burel G. Blind recovery of k/n rate convolutional encoders in a noisy environment. *EURASIP Journal on Wireless Communications and Networking*. 2011. Article 168. https://doi.org/10.1186/1687-1499-2011-168.
5. Nikolaev S.N., Romanov A.N. Method for recognition of parameters of error-correcting block-cyclic codes by a generator polynomial. *Cybernetics and Systems Analysis*. 2021. Vol. 57, N 1. P. 146–154. https://doi.org/10.1007/s10559-021-00338-w.
6. Han S., Zhang M., Li X. A blind identification method of self-synchronous scramblers based on optimization of established cost function. *Journal of Electronics & Information Technology.* 2018. Vol. 40, Iss. 8. P. 1971–1977. https://doi.org/10.11999/JEIT171026.

7. Chuklin A., Schuth A., Zhou K., De Rijke M. A comparative analysis of interleaving methods for aggregated search. *ACM Transactions on Information Systems*. 2015. Vol. 33, N 2. Article 5. https://doi.org/10.1145/2668120.
8. Ніколаєв С.М., Нищук А.М. Статистичні ознаки мовних кодеків в комплексах спеціального призначення. Проблеми кібербезпеки інформаційно-телекомунікаційних систем. *Тези доповідей II наук.-практ. конф*. (23–24 березня 2017, Київ, Україна). Київ, 2017. С. 146–147.
9. Сизов А.С., Сидоренко С.М., Сидоренко С.С., Москалев А.А. Структурные свойства произвольных дискретных последовательностей. Часть 2. Характеристика типов периодичностей произвольных дискретных последовательностей. *Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение*. 2012. № 2, Ч. 3. С. 214–218.
10. Табунщик Г.В., Кудерметов Р.К., Каплієнко Т.І. Інженерія якості програмного забезпечення. Запоріжжя: Дике Поле, 2016. 176 с.
11. Hamming R.W. Error detecting and error correcting codes. *The Bell System Technical Journal*. 1950. Vol. 29, Iss. 2. P. 147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x.
12. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов. *Доклады Академии Наук СССР*. 1965. Т. 163, № 4. С. 845–848.
13. Wagner R.A., Fischer M.J. The string-to-string correction problem. *Journal of the ACM*. 1974. Vol. 21, Iss. 1. P.168–173. https://doi.org/10.1145/321796.321811.
14. Damerau F.J. A technique for computer detection and correction of spelling errors. *Communications of the ACM*. 1964. Vol. 7, N 3. P. 171–176. https://doi.org/10.1145/363958.363994.
15. Давыдова Ю.В. Модель ошибок для нечеткого текстового поиска в задаче мониторинга виртуальных социальных сетей для обеспечения информационно-психологической безопасности личности. *Современные информационные технологии и ИТ-образование*. 2017. Т. 13, № 3. С. 72–82. https://doi.org/10.25559/SITITO.2017.3.460.
16. Jaro M.A. Advances in record linkage methodology as applied to the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*. 1989. Vol. 84, Iss. 406. P. 414–420. https://doi.org/10.2307/2289924.
17. Winkler W.E. String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. *Proceedings of the Survey Research Methods Section*. American Statistical Association. 1990. P. 354–359.
18. Needleman S.B., Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. 1970. Vol. 48, Iss. 3. P. 443–453. https://doi.org/10.1016/0022-2836(70)90057-4.
19. Smith T.F., Waterman M.S. Identification of common molecular subsequences. *Journal of Molecular Biology*. 1981. Vol. 147, Iss. 1. P. 195–197. https://doi.org/10.1016/0022-2836(81)90087-5.
20. Gotoh O. An improved algorithm for matching biological sequences. *Journal of Molecular Biology.* 1982. Vol. 162, Iss. 3. P. 705–708. https://doi.org/10.1016/0022-2836(82)90398-9.
21. Jaccard P. The distribution of the flora in the alpine zone. *New Phytologist*. 1912. Vol. 11, N 2. P. 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.
22. Berger A., Guda S. Threshold optimization for F measure of macro-averaged precision and recall. *Pattern Recognition.* 2020. Vol. 102. Article 107250. https://doi.org/10.1016/j.patcog.2020.107250.
23. Hand D., Christen P., Kirielle N. F*: an interpretable transformation of the F-measure. *Machine Learning*. 2021. Vol. 110, Iss. 3. P. 451–456. https://doi.org/10.1007/s10994-021-05964-1.

A. Nyshchuk, S. Nikolaev, O. Romanov

TECHNIQUES FOR ANALYZING BITSTREAMS BASED ON THE USE OF DAMERAU–LEVENSHTEIN DISTANCE AND OTHER METRICS

Abstract. The analysis of the features of various distances (metrics) in solving the problems of comparing bitstreams was carried out. The application of the proposed method for solving the problem of identification of bitstreams of digital speech coding is considered as an example. The greatest effectiveness of the Damerau–Levenshtein metric in this case is shown.

Keywords: bitstream, distance, metric, digital speech encoding.

Надійшла до редакції 20.04.2023