Одеський національний університет імені І. І. Мечникова
Факультет математики, фізики та інформаційних технологій
Кафедра оптимального керування і економічної кібернетики

Кваліфікаційна робота

на здобуття ступеня вищої освіти «бакалавр»

«Ідентифікація звуків на основі перетворення Гільберта-Хуанга»

«Sound identification based on the Hilbert-Huang transform»

Виконала: студентка денної форми навчання
спеціальності 113 Прикладна математика
Марія БЄРЄСТОВА

Керівник: канд. техн. наук, доц. Володимир МОРОЗ
Рецензент: канд. фіз.-мат. наук, доц. Віктор ВЕРБІЦЬКИЙ

Рекомендовано до захисту:

Протокол засідання кафедри

№ ___ від _____ 2025 р.

Завідувач кафедри

_____ _____

Захищено на засіданні ЕК № _____

Протокол № ___ від _____ 2025 р.

Оцінка _____ / _____ / _____

Голова ЕК

_____ _____

Одеса — 2025 р.

ЗМІСТ

ВСТУП

Традиційні методи аналізу сигналів ґрунтуються на припущеннях лінійності та стаціонарності. У XX сторіччі були розроблені нові методи часо-частотного аналізу сигналів, наприклад, вейвлетний аналіз та розподіл Вігнера-Вілля, які підходять для нелінійних, або нестаціонарних даних. Для нелінійних даних, які є стаціонарними, розроблено методи аналізу часових рядів. Однак більшість реальних систем, як природних, так і створених людиною, генерують дані, які одночасно нелінійні й нестаціонарні. Аналіз таких даних є складним завданням, оскільки стандартні математичні підходи, що використовують заздалегідь задані базові функції, виявляються неефективними. Для розв'язання цієї проблеми потрібна адаптивна основа - базис, що залежить від самих даних. Адаптивний базис повинен залежати від даних та визначатися a posteriori.[1]

Через це, часо-частотний аналіз та побудова адаптивних базисів для сигналу є актуальною задачею.

Перетворення Гільберта — Хуанга частково розв'язує ці проблеми та дозволяє ефективно аналізувати нелінійні та нестаціонарні дані, особливо для представлення часу, частоти та енергії. Метод був перевірений емпірично і дав результати, значно точніші, ніж традиційні методи. [2]

Перетворення Гільберта - Хуанга (англ. HHT) - це перетворення, яке є розкладанням сигналу на емпіричні моди, з подальшим застосуванням до отриманих компонентів розкладання перетворення Гільберта. Потужність та ефективність HHT в аналізі даних були продемонстровані його успішним застосуванням до багатьох важливих проблем, що охоплюють інженерні (аналіз вібрацій у будівельних конструкціях) [3], біомедичні (аналіз електрокардіограм або електроенцефалограм) [4], фінансові (виявлення ринкових трендів та волатильності) [5] та геофізичні (аналіз землетрусів та океанічних хвиль)[6], [7] дані.

Метою роботи є дослідження можливостей перетворення Гільберта—Хуанга для ідентифікації звуків, аналізу їхніх часових та частотних характеристик у порівнянні з іншими методами обробки сигналів.

Об'єкт дослідження – часо-частотний аналіз аудіосигналів, методи

представлення обробки та ідентифікації аудіосигналів.

Предмет дослідження — ефективне розпізнавання звуків на основі нейромережевого підходу та порівняльний аналіз з класичними методами.

Для досягнення поставленої мети роботи були розв'язані наступні задачі:

1) аналіз методів часо-частотного аналізу;
2) реалізація емпіричного модового розкладання (EMD) та отримання внутрішніх модових функцій залежно від локальних характеристик сигналу;
3) модифікація перетворення Гільберта–Хуанга (HHT), що полягає у вдосконаленні алгоритму з урахуванням специфічних особливостей аудіоданих.

РОЗДІЛ 1

ПОСТАНОВКА ЗАДАЧІ

1.1 Актуальність проблеми

Метод перетворення Гільберта–Хуанга (HHT), який поєднує емпіричне модове розкладання (EMD) із перетворенням Гільберта, забезпечує адаптивний підхід до аналізу сигналів. HHT дозволяє ефективно виділити часово-частотні характеристики навіть у складних шумових умовах, оскільки базис будується на основі самих даних. Однак, незважаючи на успішне застосування HHT у біомедицині, фінансах та інженерії, його адаптація до специфіки аудіосигналів акустичного моніторингу, зокрема для виявлення дронів, потребує поглибленого дослідження, розробки відповідних модифікацій та порівняння з класичними підходами — такими як MFCC-SVM чи нейромережеві архітектури CNN.

1.2 Основні терміни та поняття

Означення 1.2.1. Нестаціонарний сигнал - сигнал, характеристики якого змінюються у часі.

Означення 1.2.2. Перетворення Гільберта–Хуанга (HHT) - метод адаптивного часово-частотного аналізу, який складається з емпіричного модового розкладання (EMD) та подальшого Гільберт-перетворення.

Означення 1.2.3. Емпіричне модове розкладання (EMD) - алгоритм, що розкладає сигнал на внутрішні модові функції (IMF) на основі локальних екстремумів сигналу.

Означення 1.2.4. Внутрішня модова функція (IMF) - складова сигналу з фізично осмисленим частотним змістом, яка задовольняє певні математичні критерії (наприклад, кількість екстремумів і нулів не повинна відрізнятись більш ніж на одиницю).

Означення 1.2.5. MFCC (Mel-Frequency Cepstral Coefficients) - мел-частотні кепстральні коефіцієнти — ознаки, які широко використовуються для аналізу мови та звуків у системах розпізнавання.

Означення 1.2.6. CNN (Convolutional Neural Network) - згорткова нейронна мережа — клас глибоких нейронних мереж, ефективний у задачах класифікації зображень, спектрограм та аудіосигналів.

1.3 Огляд сучасних підходів

Відомі методи обробки аудіосигналів включають класичні підходи, такі як MFCC (мел-частотні кепстральні коефіцієнти) у поєднанні з методами машинного навчання; нейромережеві методи, такі як згорткові нейронні мережі (CNN), рекурентні (RNN/CRNN), трансформери для обробки спектрограм; методи адаптивного аналізу (вейвлети, віконне Фур'є-перетворення, а також HHT).

РОЗДІЛ 2

ОГЛЯД НАЯВНИХ РІШЕНЬ ДЛЯ ІДЕНТИФІКАЦІЇ ЗВУКІВ

2.1 Furier transform

Фур'є-перетворення (FT) та швидке Фур'є-перетворення (FFT) – широко використовуються для спектрального аналізу звукових сигналів. Фур'є-перетворення дозволяє представити сигнал у вигляді сукупності синусоїдальних компонент, що робить його зручним для подальшої обробки.

FFT - обчислювально ефективніша версія. Обчислювальна складність O(NlogN). FFT використовує властивості симетрії і періодичності, розбиваючи DFT на менші частини, які можуть бути обчислені рекурсивно. Наукові дослідження підтверджують ефективність FFT у класифікації звукових сигналів; FFT є основою для MFCC, DFT.

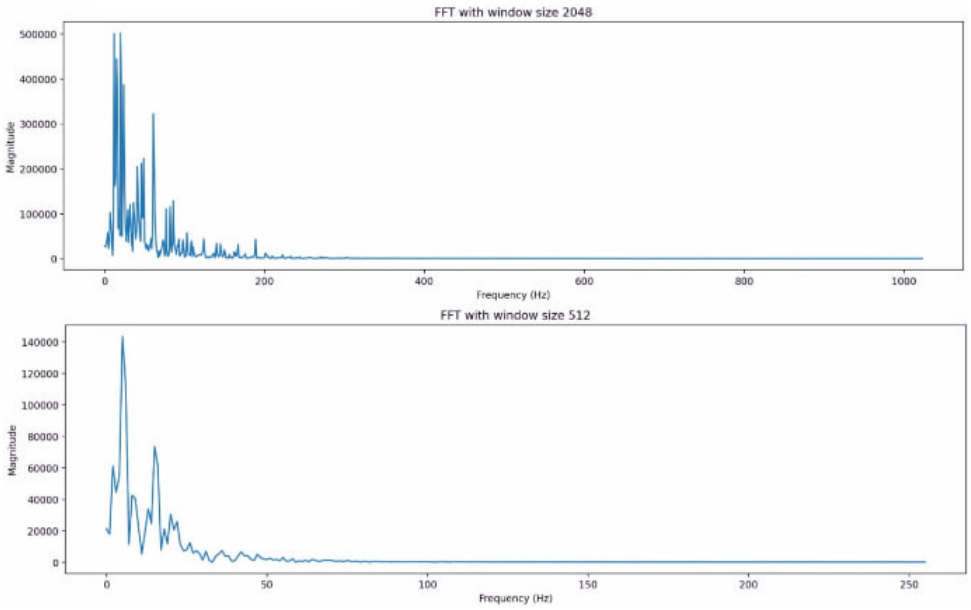

Рис. 2.1. Приклади FFT з різним розміром вікна.

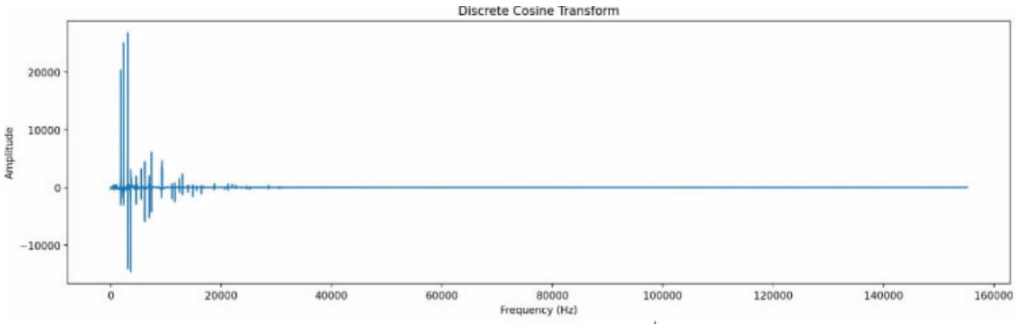

Рис. 2.2. Дискретне косинусне перетворення

2.1.1 MFCC(Mel-frequency cepstral coefficients)

MFCC (Mel-frequency cepstral coefficients):

$$m(f) = 2595\ \log_{10}\Big(1 + \tfrac{f}{700}\Big) \quad [\text{mel}], \tag{2.1}$$

де f — частота в герцах. MFCC подають спектр сигналу в цьому перцептивно релевантному масштабі й тому стали де-факто стандартом для систем розпізнавання мовлення.

Алгоритм обчислення:

Нехай $x[n]$ — цифровий сигнал із частотою дискретизації f_s. Спочатку, сигнал ділять на перекривані кадри довжиною N семплів (типово 20–32 мс), множать на вікно $w[n]$ (частіше за все Хеммінга). До кожного кадру застосовують ДПФ: $X[k] = \sum_{n=0}^{N-1} x[n]\, w[n]\, e^{-j\frac{2\pi}{N}kn}$. Беруть лише модуль $|X[k]|$. Модульний спектр пропускають через M трикутних фільтрів, розміщених рівномірно на mel-шкалі:

$$E_m = \sum_{k=f_{m-1}}^{f_{m+1}} |X[k]|^2\, H_m[k], \qquad m = 1, \ldots, M.$$

Логарифм енергій: $\tilde{E}_m = \log E_m$. Дискретне косинусне перетворення (DCT):

$$c_n = \sum_{m=1}^{M} \tilde{E}_m\, \cos\Big[\frac{\pi n}{M}\,\big(m - \tfrac{1}{2}\big)\Big], \qquad n = 0, \ldots, L-1. \tag{2.2}$$

Зазвичай беруть $L = 12$–13 найнижчих коефіцієнтів. Опційно - ліфтинг і

дельти. Використовують кепстральне згладжування (*liftering*) та додають похідні Δc_n і $\Delta^2 c_n$ для моделювання динаміки.

Щодо недоліків MFCC, то логарифм енергій підсилює низькоенергетичний шум, що погіршує роботу ASR у несприятливих акустичних умовах. Також використовується лише амплітуда спектра, тоді як фаза може містити корисні ознаки. Також припущується стаціонарність всередині кадру. Швидкі переходні процеси (наприклад вибухові приголосні) описуються неточно, якщо їхня тривалість $<$ довжини кадру. Мел-шкала оптимізована під слухову модель мовлення; для не-мовних сигналів (музика, технічні шуми) може бути далекою від оптимальної.

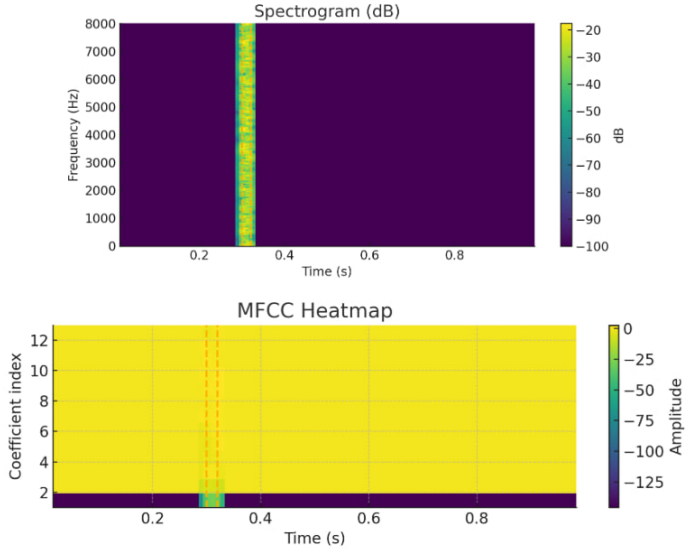

Рис. 2.3. Приклад MFCC на синтетичному "вибуховому"сплеску

2.1.2 DFT та warped-DFT

Слід також описати дискретне перетворення Фур'є (DFT), що перетворює кінцеву послідовність дискретних зразків у комплексний спектр частотних компонент. Обчислювальна складність - зростає квадратично зі збільшенням кількості зразків.

Для скінченної послідовності $x[n] \in \mathbb{C}$, $n = 0, \ldots, N-1$, *дискретне перетворення Фур'є* визначається формулою

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-\mathrm{j}\frac{2\pi}{N}kn}, \qquad k = 0, \ldots, N-1. \tag{2.3}$$

Інверсне перетворення (IDFT) відновлює $x[n]$ із спектра $X[k]$:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k]\, e^{\,\mathrm{j}\frac{2\pi}{N}kn}, \qquad n = 0, \ldots, N-1. \tag{2.4}$$

Ключові властивості:

(i) **Лінійність:** $\mathcal{F}\{ax_1[n] + bx_2[n]\} = aX_1[k] + bX_2[k]$.

(ii) **Парсеваль:**
$$\sum_{n=0}^{N-1} |x[n]|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X[k]|^2.$$

(iii) **Циклічне згортання:** якщо $z[n] = x[n] \circledast_N y[n]$, то $Z[k] = X[k]\, Y[k]$.

(iv) **Часове зсування:** $x[(n - n_0) \bmod N] \iff e^{-\,\mathrm{j}\frac{2\pi}{N}kn_0} X[k]$.

(v) **Частотне зсування:** $e^{\,\mathrm{j}\frac{2\pi}{N}k_0 n} x[n] \iff X[(k - k_0) \bmod N]$.

DFT можна розглядати як рівномірне дискретизування *дискретного часового Фур'є-перетворення* (DTFT) $X(e^{\mathrm{j}\omega})$:

$$X[k] = X\big(e^{\mathrm{j}\omega_k}\big), \quad \omega_k = \tfrac{2\pi k}{N}.$$

Це еквівалентно припущенню, що $x[n]$ є періодичним з періодом N.

Наївна реалізація формули (2.3) потребує $\mathcal{O}(N^2)$ комплексних множень. *Швидке перетворення Фур'є* (FFT) знижує складність до $\mathcal{O}(N \log_2 N)$, використовуючи факторизацію матриці $\mathbf{F}_N$ за схемою «розділяй та владарюй». Завдяки цьому FFT є стандартом для практичної спектральної обробки сигналів, фільтрування методом частотного домену, модуляцій OFDM тощо.

Також існує *warped DFT* [8]. На відміну від класичного ДПФ з рівномірною частотною сіткою (2.3), *warped DFT* розміщує відліки нерівномірно, що дає змогу адаптивно деталізувати «важливі» частотні ділянки без збільшення N; краще узгоджуватися з психоакустичними або нелінійними шкалами; підвищувати точність оцінювання параметрів коротких сигналів.

Однак WDFT втрачає деякі ортогональні та конволюційні властивості ДПФ, вимагаючи чисельно стійких методів інверсії (наприклад, SVD).

2.1.3 Класифікація звуків на основі перетворення Фур'є

Автор [9] побудував систему автоматичної класифікації мажорних та мінорних акордів на основі спектральних ознак аудіосигналу. Кожен `.wav`-файл декодується, після чого на відтинку довжини N відліків обчислюється швидке перетворення Фур'є $X_F = \mathrm{FFT}\big(x[n]\big)$, а з нього формується односторонній амплітудний спектр $A(f) = \frac{2}{N}|X_F[0{:}N/2]|$ та вектор частот $f_k = k\, f_s/N$. Для придушення індустріального шуму вилучаються компоненти нижче 50 Гц; далі алгоритм `find_peaks` із міжпіковою відстанню 10 бінів і порогом 5 % від глобального максимуму виявляє локальні екстремуми, частоти яких інтерпретуються як гармоніки. Записуються мінімальна й максимальна гармоніка, їхня кількість та до 20 окремих частот, після чого формуються відносні інтервали $I_i = f_{i+1}/f_i$, інваріантні до абсолютної висоти звуку. Цей підхід надав точність близько 92 % для Random Forest та для CNN1D.

У дослідженні [10] FFT використовувалося для виявлення аномалій у промислових акустичних сигналах. Запропонована система складається з двох головних компонентів: серверної частини та мобільного застосунку для збору й надсилання аудіоданих.

Реалізовано чотири модулі екстракції: дискретне косинусне перетворення (DCT), *multiple window* швидке перетворення Фур'є (FFT), *single window* FFT та мел-частотні кепстральні коефіцієнти (MFCC). Кожен модуль обробляє початковий звуковий сигнал, формуючи вектор ознак $\mathbf{x} \in \mathbb{R}^d$.

Базовим класифікатором є k-найближчих сусідів із евклідовою метрикою ($k = 1$). Для нового вектора ознак $\mathbf{y}$ мітка l визначається як

$$l = L\big(\arg\min_i \|\mathbf{y} - \mathbf{x}_i\|_2\big), \tag{2.5}$$

де $\mathbf{x}_i$ — вектори ознак із відомими мітками, а $L(i)$ — відображення індекса i у відповідну мітку. Спершу класифікатори тренуються лише на *нормальних* звуках і працюють як детектори викидів. Після виявлення потенційної аномалії користувач підтверджує або відхиляє її, що дає змогу *донавчити* модель та додавати нові мітки. Вихід кількох класифікаторів комбінується методом зваженого голосування, де вага кожного дорівнює його точності

на тренувальній вибірці.

В іншій роботі [11] пропонуються два незалежні підходи до ідентифікації безпілотників за їхнім акустичним підписом: **(i) кореляційний аналіз спектра** та **(ii) аудіо-фінґерпринтинг**.

Вхідний сигнал дискретизується , перетворюється в частотну область швидким перетворенням Фур'є і обмежується інформативною смугою $f > 5$ кГц, яка відповідає шуму пропелерів.

Серед випробуваних статистичних показників автори зосередилися на коефіцієнті Пірсона

$$r = \frac{\sum_{i=1}^{n} \big(x_i - \overline{x}\big)\big(y_i - \overline{y}\big)}{\sqrt{\sum_{i=1}^{n} \big(x_i - \overline{x}\big)^2} \sqrt{\sum_{i=1}^{n} \big(y_i - \overline{y}\big)^2}}, \tag{2.1}$$

та *нормалізованій максимальній кореляції* (NMC)

$$\mathrm{NMC}(x,y) = \frac{\max_{\tau} \big|\mathrm{xcorr}_{\tau}(x,y)\big|}{\|x\|_2 \, \|y\|_2}, \tag{2.2}$$

яка, на відміну від (2.1), інваріантна до масштабних відмінностей сигналів.

Для квадрокоптера самокореляційний пік склав 2728 ум. од., тоді як найближчий хибний збіг (фен для волосся) — 443 ум. од. Відповідно поріг виявлення було встановлено так, щоб забезпечити $< 5\%$ хибних спрацювань на робочій відстані до 1,5–2 м (мікрофон Behringer C1-U).

Для кожного секційного кадру формується спектрограма, у часово-частотній площині виділяються *стійкі точки*, а з найближчої пари піків (f_1,t_1) та (f_2,t_2) обчислюється хеш $H = \langle f_1, f_2, \Delta t \rangle$, де $\Delta t = t_2 - t_1$. У середньому виходить приблизно 300 хешів на секунду сигналу.

Подібність двох сигналів визначається евклідовою відстанню між множинами їх хешів:

$$\|F(X) - F(Y)\| < T \implies X \equiv Y, \tag{2.3}$$

де T — емпірично підібраний поріг.

За чистих умов середня схожість хешів між різними квадрокоптерами досягала 78,9%, тоді як для побутових шумів не перевищувала 15%. Метод виявився стійкішим до адитивного шуму та стиснення (MP3, AAC) і працював на відстанях до 3 м.

Робота [11] демонструє, що:

а) вузькосмугове кореляційне зіставлення спектрів забезпечує високу точність за доброякісного сигналу, проте швидко деградує при низькому SNR;
б) Shazam-подібний фінґерпринтинг є робастнішим до шуму й істотно скорочує вимоги до памяті;
в) комбінована схема «fingerprint →NMC-верифікація» потенційно підвищує дальність виявлення та зменшує кількість хибних спрацювань.

У роботі [12] досліджується здатність warped DFT давати кращі ознаки для розпізнавання без додаткових прийомів шумозахисту, ніж класичні MFCC та PLP (Perceptual Linear Prediction). Автори дійшли до висновку, що WDFT-LP (MFCC, сформовані зі спектра, змодельованого all-pole (LP) після WDFT) забезпечує стійкіші до шуму та каналів спектральні ознаки, ніж стандартні MFCC і PLP. Помилка розпізнавання знижується до 33% (clean training) і 22% (multistyle).

2.2 Wavelet transform

Вейвлет-перетворення дозволяє аналізувати сигнали на різних частотах з різною роздільною здатністю, що особливо корисно для динамічних звукових сигналів. Це перетворення ефективне для обробки як високочастотних, так і низькочастотних компонентів сигналу. Нехай $f \in L^2(\mathbb{R})$ та *материнська вейвлет-функція* ψ задовольняє *умову допустимості*

$$C_\psi := \int_0^\infty \frac{|\widehat{\psi}(\omega)|^2}{\omega}\, d\omega < \infty.$$

Тоді **неперервне вейвлет-перетворення** визначається так

$$W_\psi f(a,b) \;=\; \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t)\, \overline{\psi\left(\frac{t-b}{a}\right)}\, dt, \qquad a \in \mathbb{R} \setminus \{0\},\, b \in \mathbb{R}.$$

Інверсна формула відновлення

$$f(t) = \frac{1}{C_\psi} \int_0^{\infty} \int_{-\infty}^{\infty} W_\psi f(a,b)\, \frac{1}{|a|^{3/2}}\, \psi\left(\frac{t-b}{a}\right)\, db\, da,$$

забезпечує *ізометрію* $L^2(\mathbb{R}) \to L^2(\mathbb{R}^+ \times \mathbb{R})$.

Дискретне вейвлет-перетворення (DWT) аналізує сигнал як в часі, так і в частоті. Він використовує вейвлети - функції, які дозволяють отримати детальні та апроксимуючі коефіцієнти.

Дискретне вейвлетне перетворення (DWT) використовує дискретні значення для масштабів і зсувів, зазвичай на основі степенів двійки. Це дозволяє ефективніше обробляти і зберігати дані. Результатом є набір вейвлетних коефіцієнтів, які представляють сигнал на різних масштабах і позиціях. Перевагою є ефективність у зберіганні даних і обчисленнях, зручність для сигналів обмеженої довжини та практичних застосувань. Але може втрачати інформацію про сигнал через обмежену кількість масштабів і зсувів.

Фіксуємо *дискретну решітку* масштабу-зсуву

$$a_j = 2^{-j}, \quad b_{j,k} = k\, 2^{-j}, \qquad j,k \in \mathbb{Z}.$$

Дискретні вейвлети

$$\psi_{j,k}(t) = 2^{j/2}\, \psi(2^j t - k)$$

утворюють ортонормований базис у $L^2(\mathbb{R})$ (за наявності *мульти-роздільового аналізу*, MRA). Тоді

$$W_\psi f[j,k] \;=\; \langle f, \psi_{j,k} \rangle = \int_{-\infty}^{\infty} f(t)\, \psi_{j,k}(t)\, dt$$

є коефіцієнтами DWT. Функцію відновлюють скінченною сумою

$$f(t) = \sum_{j,k} W_\psi f[j,k]\ \psi_{j,k}(t).$$

Алгоритм Маллата (фільтр-банк)

Нехай φ — *функція масштабування*, $h[\cdot]$ та $g[\cdot]$ — *квадратурно-дзеркальні фільтри* (низько- і високочастотний). Для *наближених* $a_j[\cdot]$ та *детальних* $d_j[\cdot]$ коефіцієнтів:

$$a_{j+1}[n] = \sum_k h[k-2n]\, a_j[k],$$
$$d_{j+1}[n] = \sum_k g[k-2n]\, a_j[k].$$

Зворотна стадія (синтез) використовує ті ж фільтри:

$$a_j[k] \ = \ \sum_n h[k-2n]\, a_{j+1}[n] \ + \ \sum_n g[k-2n]\, d_{j+1}[n].$$

Ключові властивості DWT

- **Ортонормованість**: $\langle \psi_{j,k}, \psi_{j',k'} \rangle = \delta_{j,j'}\delta_{k,k'}$.
- **Стиснення**: більшість $d_j[k]$ часто близькі до нуля $\Rightarrow$ ефективне кодування.
- **Локальність**: одночасно хороша часово-частотна роздільна здатність.
- **Швидкість**: алгоритм Маллата виконується за $O(N)$ операцій для сигналу довжини N.

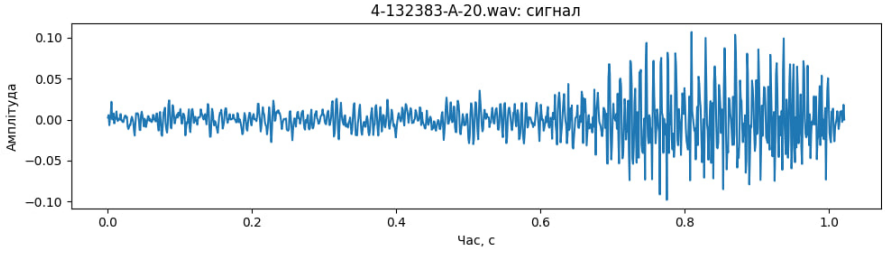

Рис. 2.4. Сигнал

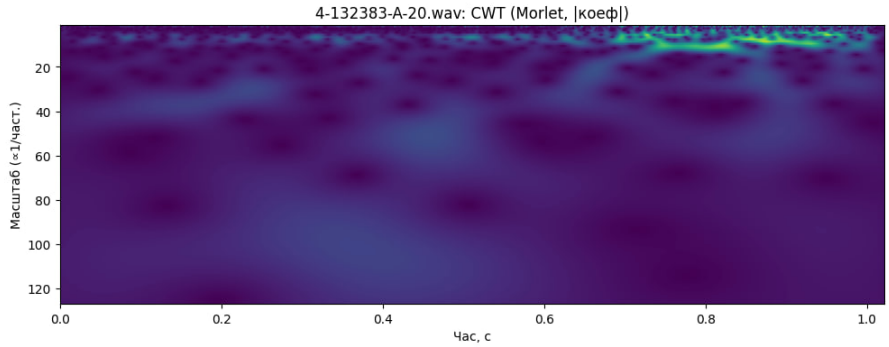

Рис. 2.5. Приклад CWT

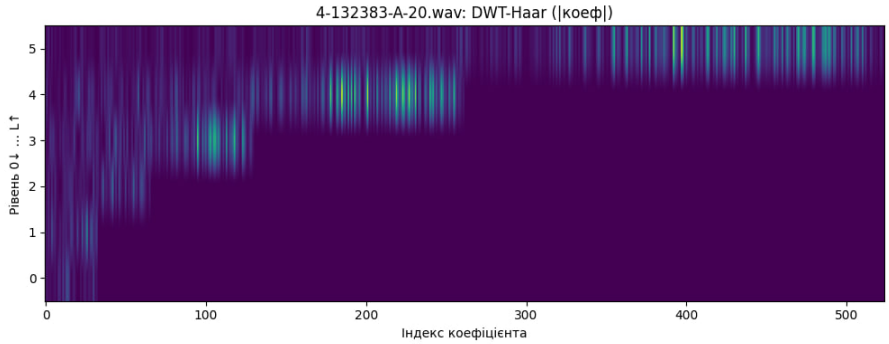

Рис. 2.6. Приклад DWT

Дослідження [13] спрямоване на підвищення стійкості ізольованого розпізнавання слів за рахунок заміни стаціонарної дискретної Фур'є-бази в класичному pipeline MFCC на адаптивне багаторівневе хвильове пакетне перетворення (Wavelet-Packet Transform, WPT). Автори застосовують материнську хвильку Daub-1, виконують декомпозицію до четвертого рівня і залишають лише піддіапазони найбільшої енергії (1,0), (2,2), (3,2), (4,0), (4,1), (4,2), (4,8) та (4,12), що істотно знижує розмірність ознак без втрати інформативності.

Сигнал спершу піддають пре-емфазі фільтром

$$y[n] = x[n] - 0.95\,x[n-1],$$

після чого сегментують у 25-мс вікна з перекриттям і вікнують функцією Хеммінга. До кожного кадру застосовують WPT; з обраних піддіапазонів обчислюють логарифмічні енергії, а потім виконують дискретне косинусне перетворення, отримуючи кепстральний вектор. Оскільки реальна кількість

кадрів s варіює, вводять процедуру *frame-normalization*: для наперед заданого t вибирають кадри з індексами $\lceil i\,s/t \rceil$ $(i = 1, \ldots, t)$ або дублюють сусідні, що гарантує сталий розмір входу класифікатора.

Класифікацію реалізовано двома моделями. Це трьохшарова штучна нейронна мережа з сигмоїдальними прихованими та лінійним вихідним шаром та багатокласовий SVM.

Експерименти у MATLAB на 48 тренувальних і 12 тестових сигналів демонструють істотну перевагу хвильових ознак. «Method 1» (WPT + DCT + frame-normalization) у парі з ANN досягає точності 91.67%, тоді як класичні MFCC із тією ж ANN — лише 66.67%. Для SVM контраст ще виразніший: 73% проти 50%. «Method 2», де ознаковий вектор утворюється простим конкатенуванням коефіцієнтів усіх вибраних піддіапазонів і скорочується PCA, показує проміжні результати (66.67% з ANN, 61.67% з SVM).

2.3 Perceptual Linear Prediction (PLP)

PLP [14] — це спосіб перетворити спектр аудіосигналу в компактний набір коефіцієнтів, максимально наближений до людського сприйняття гучності та частоти. RASTA-PLP (**R**elative **A**uditory **S**c**T**ionary **A**pproximation) додає до PLP фільтрацію по часу, щоб подавити повільні (станційні) та дуже швидкі (імпульсні) зміни, які часто є шумом.

Алгоритм PLP:

1) **Перетворення в частотну область:** $P(f) = |X(f)|^2$ — потужнісний спектр з віконованого STFT-кадру.
2) **Критичні смуги (Bark):**

$$B(f) = 13\,\arctan\big(0.00076\,f\big) + 3.5\,\arctan\big[(f/7500)^2\big],$$

   після чого $P(f)$ згортається з трикутним фільтробанком шириною 1 Bark.
3) **Моделювання рівної гучності:**

$$\tilde{P}(f) = P(f)\,E(f), \qquad E(f) = \frac{(f^2 + 56.8 \cdot 10^6) f^4}{(f^2 + 6.3 \cdot 10^6)^2 (f^2 + 0.38 \cdot 10^9)},$$

де $E(f)$ — формула ISO 226 для кривої рівної гучності (у спрощеному вигляді).

4) **Нелінійна компресія гучності:** $S(f) = \tilde{P}(f)^{1/3}$ (кубічний корінь апроксимує закон Стівенса [15]).
5) **Повернення у час → LP-аналіз:**
   $s[n] = \mathcal{F}^{-1}\{S(f)\}$ → автокореляція → `levinson-durbin` → LP-коефіцієнти a_k.
   Зазвичай лишають $p = 8$–16 коефіцієнтів і перетворюють їх у кепстральні c_m для стабільності:

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}.$$

RASTA-фільтрація Для кожної критичної смуги сигнал проходить фіксований IIR-фільтр

$$H(z) = \frac{0.1(1 - z^{-2})}{1 - 0.98\, z^{-1}} \quad \Longleftrightarrow \quad h[n] = 0.1\big(\delta[n] - \delta[n-2]\big) + 0.98\, h[n-1],$$

що утворює «полос-пропуск до 1 Гц» у лог-амплітудному домені: *повільні зміни (наприклад, рівень шуму) та миттєві сплески (клаци) обтинаються, залишаючи інформативну середньочастотну динаміку.*

Переваги.

- Ближчий до психоакустики, ніж MFCC: критичні смуги й еквалізація голосності.
- Менша чутливість до зміни мікрофонів та каналів (особливо RASTA-PLP).
- Підтримує LP-параметри (*formant-tracking*, *speaker ID*) й статистичні моделі (HMM-GMM, GMM-UBM).

Недоліки.

- Складніший, ніж MFCC, та має більше гіперпараметрів (порядок LP, вікно, константи фільтра).

- Cube-root і RASTA-фільтр можуть спотворювати фрикативні звуки, якщо вирізати занадто вузьке вікно.

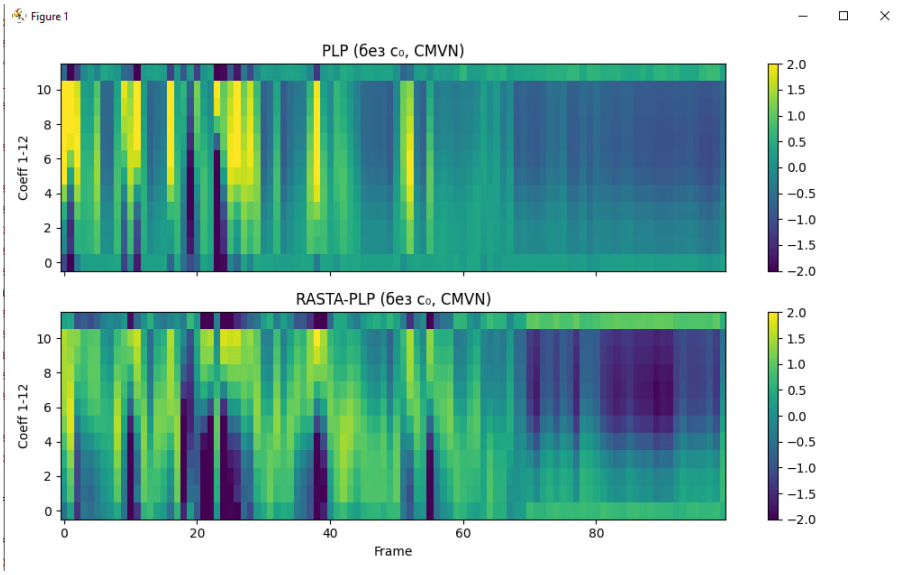

Рис. 2.7. Приклад PLP, RASTA-PLP

РОЗДІЛ 3

ІДЕНТИФІКАЦІЯ ЗВУКІВ НА ОСНОВІ HHT

3.1 Загальний опис алгоритму

Перетворення Гільберта — Хуанга запропоновано Норденом Хуангом у 1995 році під час його роботи у NASA для вивчення поверхневих хвиль тайфунів (явища, коли високочастотні хвилі з коротким розгоном переходять у низькочастотні хвилі з довгим розгоном). За допомогою методу EMD-HSA було виявлено, що еволюція хвиль була не безперервною, а різкою, дискретною та локальною. Спочатку алгоритм називався «EMD-HSA» - метод емпіричної модової декомпозиції (англ. EMD) нелінійних і нестаціонарних процесів з наступним спектральним Гільбертовим аналізом (англ. HSA). У наступні роки, у міру розширення застосування EMD-HSA для інших галузей науки й техніки, замість терміна EMD-HSA було прийнято коротший термін перетворення HHT. Алгоритм HHT запатентовано NASA.

3.1.1 Метод емпіричного розкладання мод

Отже, перший етап алгоритму - застосування методу емпіричної модової декомпозиції. Метод емпіричного розкладання мод (Empirical Mode Decomposition, EMD) був запропонований Хуангом та співавторами для аналізу даних, отриманих з нестаціонарних та нелінійних процесів. Цей підхід передбачає, що будь-які дані складаються з різних простих "внутрішніх режимів коливань званих Intrinsic Mode Functions (IMF). Кожна наступна IMF завжди буде мати нижчу частоту, ніж попередня: перша IMF завжди містить компоненти з високою частотою, а остання – лише одну частоту (тобто монотонну функцію). Основні властивості: 1. У межах всього набору даних кількість екстремумів і число нульових перетинів має збігатися або відрізнятися максимум на одиницю. 2. У будь-якій точці середнє значення огинаючих, побудованих за локальними максимумами та мінімумами, має дорівнювати нулю. 3. Кожна IMF є більш узагальненим варіантом гармо-

нічної функції, зі змінними амплітудою і частотою, що залежать від часу. [16]

Процес розкладу включає такі етапи:

1) **Визначення локальних екстремумів:** Знаходяться всі локальні максимуми та мінімуми даних. Будуються огинаючі: верхня — з'єднує максимуми, і нижня — з'єднує мінімуми, за допомогою кубічного сплайна.
2) **Обчислення середньої огинаючої:** Середня огинальна позначається як m_1, а перший компонент обчислюється за формулою:

$$h_1 = x(t) - m_1, \quad x(t) \text{ — початкові дані}$$

3) **Процес "просіювання" (sifting):** Щоб h_1 стало IMF, процедура повторюється кілька разів:

$$h_{1k} = h_{1(k-1)} - m_{1k},$$

доки не буде виконано критерій зупинки (наприклад, кількість екстремумів і перетинів стабілізується). Підсумкова функція:

$$c_1 = h_{1k},$$

де c_1 — перша IMF.
4) **Видалення виділеного компонента:** Після знаходження c_1 обчислюється залишок r_1:

$$r_1 = x(t) - c_1.$$

Залишок r_1 знову використовується як нові дані для виділення наступної IMF:

$$r_2, r_3, \ldots, r_n = r_{n-1} - c_n.$$

5) **Кінцевий результат розкладу:** Вихідні дані можна представити як суму всіх виділених IMF та залишку:

$$x(t) = \sum_{j=1}^{n} c_j + r_n.$$

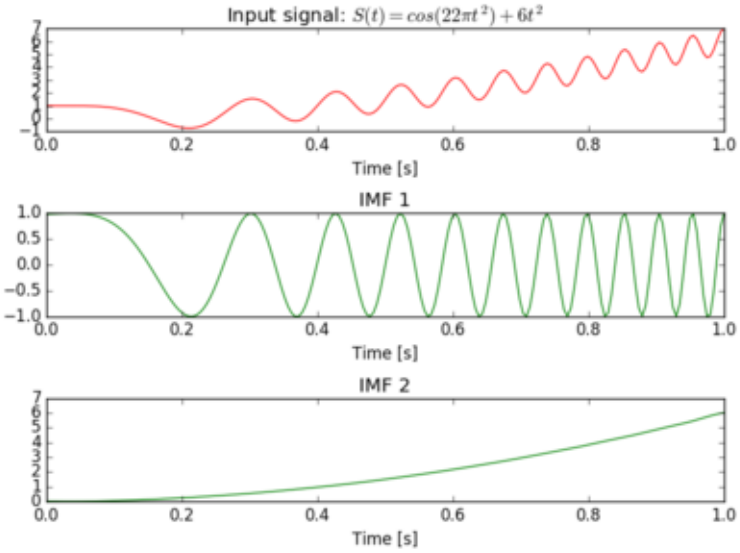

Рис. 3.1. Візуалізація кроків алгоритму

3.1.2 Критерії зупинки

Для завершення процесу розкладу застосовуються такі критерії:

- Залишкова функція r_n стає настільки малою, що її можна вважати відсутньою.
- Залишок r_n стає монотонною функцією, з якої більше неможливо виділити IMF.

EMD розбиває сигнал на набір коливальних компонент (IMF — Intrinsic Mode Functions) без використання фіксованих базових функцій, як у інших методах (наприклад, перетворення Фур'є або вейвлет-перетворення). Це важливо для аналізу нелінійних і нестаціонарних сигналів. Кожна IMF містить лише одну переважну частоту, яка може змінюватися з часом.

Рис. 3.2. Схема алгоритму

3.1.3 Гільбертовий спектральний аналіз (HSA)

HSA — аналіз частотно-часових характеристик сигналу. Він дозволяє обчислити аналітичний сигнал $z(t)$, через який визначається миттєва частота. Перетворення Гільберта $H[x(t)]$ для функції $x(t)$ із класу L_p визначається як:

$$H[x(t)] = \frac{1}{\pi}\,\mathrm{P.V.}\int_{-\infty}^{\infty}\frac{x(\tau)}{t-\tau}\,d\tau,$$

де P.V. означає головне значення (principal value) сингулярного інтегралу.

На основі перетворення Гільберта аналітичний сигнал визначається як:

$$z(t) = x(t) + iy(t) = a(t)e^{i\theta(t)},$$

де:

- $a(t) = \sqrt{x^2(t) + y^2(t)}$ — миттєва амплітуда,
- $\theta(t) = \arctan\left(\frac{y(t)}{x(t)}\right)$ — миттєва фаза,
- $y(t)$ — комплексне спряження $x(t)$, що обчислюється через перетворення Гільберта.

Миттєва частота визначається як похідна фази:

$$\omega(t) = \frac{d\theta(t)}{dt}.$$

Перетворення Гільберта спочатку застосовувалося лише до вузькосмугових сигналів, де кількість екстремумів та перетинів нуля збігається. Фільтрація в частотній ділянці (лінійна операція) видаляє гармоніки, що спотворює хвильову форму сигналу. Пряме застосування перетворення Гільберта іноді призводить до значень частоти, які можуть бути позитивними та негативними, що створює труднощі інтерпретації. Однак, EMD дозволяє розкласти сигнал на базові компоненти (IMF), що робить перетворення Гільберта ефективним.

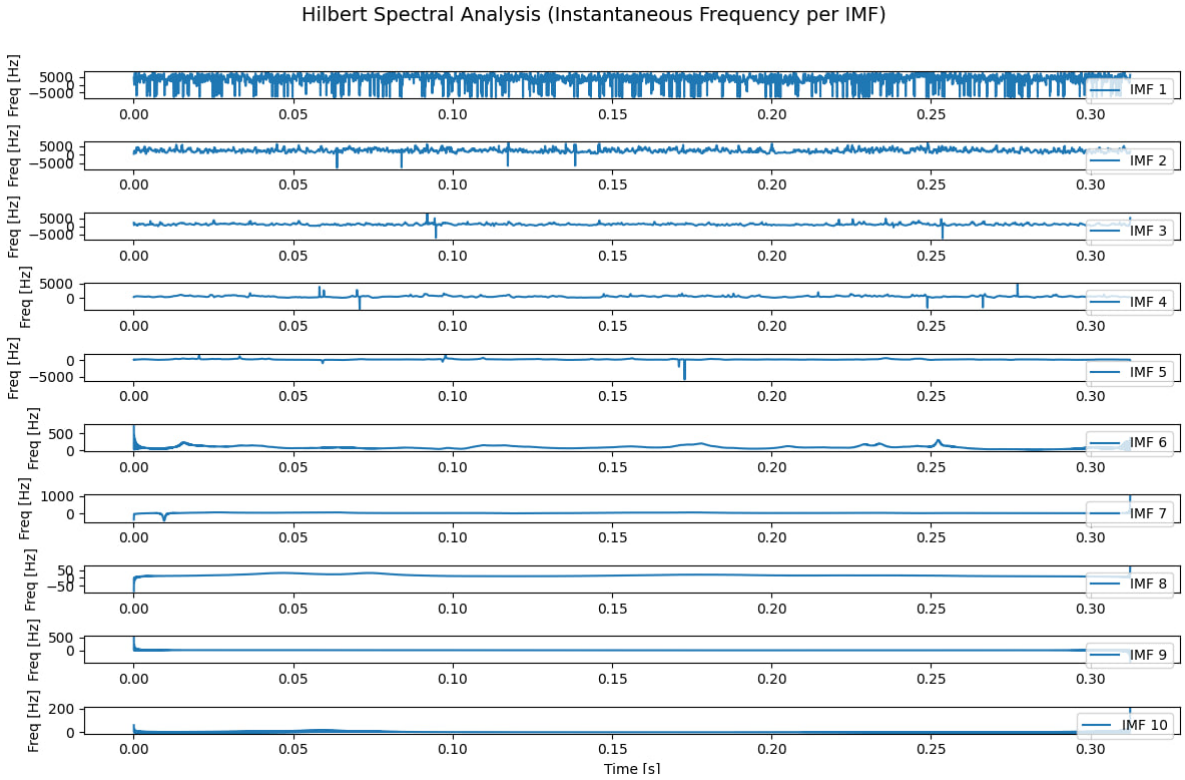

Рис. 3.3. Приклад HSA для кожної з отриманих емпіричних мод

3.2 Реалізація DCT-HHT

Детальний розв'язок можна подивитись на github. Див. Додаток А.

3.2.1 Методика обробки аудіосигналу

Для експериментальної перевірки алгоритму використовувався двокласовий корпус реальних аудіозаписів [17]: 2075 зразків класу `yes_drone` і 266 зразків класу `unknown`. Обчислення виконувалися у середовищі Python 3.12 із використанням бібліотек NumPy, SciPy, librosa та TensorFlow. Описана нижче послідовність етапів попередньої обробки й класифікації відображає типовий pipeline обробки мовно-звукових сигналів і ґрунтується на поєднанні класичних методів DSP (Digital Signal Processing) та сучасних методів машинного навчання.

3.2.2 Фільтрація сигналу

На першому етапі виконано високочастотну фільтрацію сигналу за допомогою фільтра Баттерворта 5-го порядку з граничною частотою $f_c \in \{80{,}100{,}120\}$ Гц. Фільтри Баттерворта забезпечують максимально плаский відгук у смузі пропускання (відсутність наколихувань амплітуди), що зменшує спотворення сигнального вмісту поблизу частоти зрізу. Для цифрового фільтра високих частот ступеня n із нормалізованою граничною частотою ω_c передавальна функція на j–осьовій частині може бути записана як

$$|H(j\omega)|^2 = \frac{\omega^{2n}}{\omega^{2n} + \omega_c^{2n}}\,.$$

У наших обчисленнях фільтрація виконувалась за допомогою реалізації SciPy, що забезпечує лінійно-фазову структуру і мінімальні артефакти в смузі пропускання. Головне призначення цього етапу – видалення низькочастотних шумів (наприклад, гулу) з метою покращення співвідношення сигнал/шум та уникнення спотворень у наступних кроках.

3.2.3 Сегментація та віконування (фреймінг)

Після фільтрації сигнал розбивався на перекриваючі фрейми тривалістю $L \in \{20{,}25{,}30\}$ мс при 50% перекритті, а кожен фрейм множився на вікно Хеммінга. Поділ на короткі фрейми вводиться для наближення умов

стаціонарності сигналу всередині фрейма. Зазвичай для мовних сигналів використовуються фрейми довжиною від 20 до 40 мс. Накладання вікон (50%) зменшує втрати інформації на межах фреймів та забезпечує плавність переходів.

Вікно Хеммінга застосовується для кожного фрейму з метою зменшення спектральних викидів (leakage) при перетворенні в частотну область. Його функція визначається так:

$$w[n] = 0.54 - 0.46\cos\Big(\frac{2\pi n}{N-1}\Big), \quad 0 \le n \le N-1,$$

де N — число відліків у фреймі. При цьому відлік у відфільтрованому й віконованому фреймі становить $y[n] = x[n]\,w[n]$. Використання вікна Хеммінга зменшує побічні хиби спектра (через «гладкий» перехід до нульових значень на краях), що є стандартною практикою при обробці мовних сигналів.

3.2.4 Дискретне косинусне перетворення (DCT)

Після віконування до кожного фрейма застосовувалися перетворення для вирівнювання спектра. У цьому дослідженні для кожного відокремленого фрейма застосовано дискретне косинусне перетворення (DCT):

$$X_k = \sum_{n=0}^{N-1} x[n]\cos\left(\frac{\pi}{N}(n+\tfrac{1}{2})k\right), \quad k = 0,\ldots,N-1.$$

Перетворення виконує роль спектрального еквалайзера, вирівнюючи енергетичний розподіл частотних компонентів перед подальшим аналізом. Аналогічне використання DCT описано, наприклад, при побудові мел-кепстральних коефіцієнтів для подальшої класифікації.

3.2.5 Алгоритм Гільберта–Хуанга

Наступний важливий етап — виділення амплітудної огинаючої сигналу з використанням аналізу Гільберта–Хуанга (HHT). HHT полягає у тому,

щоб отримати аналітичний сигнал із реального, а з нього — амплітудну огинаючу й миттєву фазу. Для дискретного сигналу $x[n]$ його аналітичний сигнал визначається як

$$x_a[n] = x[n] + j\,\mathcal{H}\{x[n]\},$$

де $\mathcal{H}\{x[n]\}$ — гільбертове перетворення $x[n]$. Амплітудна огинаюча поточного фрейма обчислюється як модуль комплексного аналітичного сигналу:

$$a[n] = |x_a[n]| = \sqrt{x[n]^2 + \big(\mathcal{H}\{x[n]\}\big)^2}\,.$$

Миттєва енергія сигналу представлена через квадрат амплітудної огинаючої, а розподіл цієї енергії в часі та частоті формує Гільбертовий спектр. У цій реалізації для кожного фрейму обчислено амплітудну огинаючу $a[k]$, яка характеризує локальну амплітуду сигналу в даний момент часу.

3.2.6 Мел-частотні кепстральні коефіцієнти (MFCC)

Для представлення спектральних особливостей сигналу з урахуванням сприйняття людиною використовувалися 13 MFCC. Спочатку для кожного фрейму обчислюється спектр потужності (через БПФ) і застосовується банк мел-фільтрів, що накладаються трикутними ваговими характеристиками у відповідності до мел-шкали. Після накладення фільтрів сумується енергія сигналу у кожному мел-діапазоні, потім береться логарифм енергій.

Далі до логарифмічних енергій застосовується дискретне косинусне перетворення для отримання кепстральних коефіцієнтів. Формально k-ий коефіцієнт MFCC обчислюється як

$$c_k = \sum_{m=1}^{M} \log E[m] \cos\Big[\tfrac{\pi}{M}(m - \tfrac{1}{2})k\Big], \quad k = 1, \ldots, K,$$

де $E[m]$ — енергія у m-му мел-фільтрі, M — загальна кількість фільтрів (типово 20–40), а K — число зберігаємих коефіцієнтів (у нас $K = 13$). MFCC є класичним параметричним представленням звуку для розпізнавання мовлення та звуків, оскільки вони ефективно стискають спектральну

інформацію та враховують нелінійність людського сприйняття частот. У кінці етапу аналізу з кожного фрейма отримано вектор з 13 коефіцієнтів, який потім усереднювався по всьому аудіофайлу, щоб отримати єдину ознаку для класифікації екземпляра.

3.2.7 Класифікація за допомогою багатошарового перцептрону

Отримані усереднені вектори ознак подавалися на вхід штучної нейронної мережі — багатошарового перцептрону (MLP) з одним прихованим шаром. Архітектура мережі описувалася ваговими матрицями $W_1 \in \mathbb{R}^{64\times 13}$ і $W_2 \in \mathbb{R}^{32\times 64}$, де 64 і 32 — розміри шарів. Для прихованого шару використовувалася нелінійність ReLU, що сприяє швидкій збіжності навчання і усуває проблему згасання градієнтів. Мережею було реалізовано бінарну класифікацію (клас `yes_drone` vs. `unknown`) із використанням бінарної крос-ентропійної функції втрат:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\big[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\big].$$

Як оптимізатор було обрано алгоритм Adam з початковим кроком навчання $\alpha = 10^{-3}$, батчем розміром 16 та фіксованою швидкістю навчання. Алгоритм Adam демонструє високу ефективність та стійкість у задачах великомасштабної оптимізації, зокрема завдяки адаптивному налаштуванню моментів і незначному об'єму пам'яті. 20% даних залишено для тестування моделі після навчання.

3.2.8 Оцінка моделі та пошук гіперпараметрів

Для кількісної оцінки якості класифікації використовувалися стандартні метрики: точність (accuracy), precision, recall та F_1-міра. З огляду на нерівномірність розмірів класів та інтерес до компактності невідомого класу, основним критерієм оптимізації була F_1-міра для класу `unknown`, за умови високого recall для класу `yes_drone`. Проведено градієнтний пошук

(grid-search) по значеннях параметрів f_c, L та H (усього 27 комбінацій), щоб знайти найкращу комбінацію для заданої метрики.

Для ілюстрації розподілу ознак та помилок класифікації використовувався графічний аналіз: розсіювання двох найбільш інформативних ознак на scatter-plots і confusion matrix heat map.

3.2.9 Результати

Оптимальна конфігурація (L=25 мс, H=12.5 мс, f_c=80 Гц) досягла загальної точності 0,95 на тестовому наборі. Recall дрона досяг 0.99, тоді як F1-score для невідомого класу зросла до 0.78, що перевершує базову реалізацію MFCC + SVM та наближається до точності CNN, але вимагає лише ресурсів рівня процесора (0,15 с на 3-секундний файл).

З базовою конфігурацією — тривалість кадру 25 мс, розмір стрибка 12,5 мс та відсікання високих частот 100 Гц — мережа досягла загальної точності 0,95. Клас дронів було виявлено з частотою виявлення 0,99 та точністю 0,96, що свідчить про те, що практично жоден БПЛА не був пропущений, і лише 4% позитивних тривог були хибними. Перешкодою виявився клас невідомих: його частота виявлення досягла лише 0,66, тому приблизно третина фонових звуків все ще була неправильно класифікована як дрони.

Щоб перевірити, чи був цей дисбаланс зумовлений виключно гіперпараметрами кадрування, було проведено повний пошук по сітці за 27 налаштуваннями: тривалість вікна 20, 25 та 30 мс; розміри стрибків 10, 12,5 та 15 мс; та граничні значення високих частот 80, 100 та 120 Гц. Коригування лише цих трьох факторів вже дало помітний приріст. Поєднання вікна 25 мс, стрибка 10 мс та незначного граничного значення 80 Гц підвищило показник F1 невідомого класу до 0,78, одночасно підвищивши загальну точність до 0,953. Покращення зумовлене довшим вікном, яке фіксує більше спектральної енергії справжнього фонового шуму, та м'якшим граничним значенням 80 Гц, яке зберігає низькочастотні сигнали, що розрізняють невідомі події. Найважливіше те, що метрики дронової активності залишилися практично незмінними (повністю згадані = 0,99, точність = 0,96), що

підтверджує, що нове налаштування забезпечує збалансований компроміс. Найкращий компроміс досягається при граничному значенні 80 Гц та середньому вікні 25 мс; або посилення фільтра, або вибір дуже коротких/довгих вікон систематично погіршує F1-score невідомого.

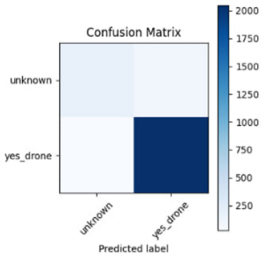

Рис. 3.4. Confusion matrix

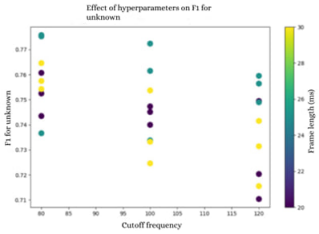

Рис. 3.5. Вплив гіперпараметрів на F1-міру

Таблиця 1

Results for frame duration = 0.020 (DCT-HHT)

Hop	**Cutoff**	**Accuracy**	**F1_unknown**	**Time**
0.0100	80	0.946604	0.752475	13.222777
0.0100	100	0.950021	0.747300	12.444511
0.0100	120	0.943614	0.720339	12.409636
0.0125	80	0.952157	0.760684	12.230076
0.0125	100	0.947458	0.739585	12.400788
0.0125	120	0.950021	0.749465	12.338219
0.0150	80	0.949594	0.743478	12.753323
0.0150	100	0.950021	0.745092	12.233622
0.0150	120	0.941478	0.710359	12.527529

Таблиця 2

Results for frame duration = 0.025 (DCT-HHT)

Hop	Cutoff	Accuracy	F1_unknown	Time
0.0100	80	0.952584	0.775758	12.371435
0.0100	100	0.952157	0.772358	12.043845
0.0100	120	0.951303	0.756410	12.277935
0.0125	80	0.947458	0.736617	12.336175
0.0125	100	0.953439	0.761488	12.487997
0.0125	120	0.951303	0.759494	12.230661
0.0150	80	0.953439	0.775258	12.188705
0.0150	100	0.945750	0.733753	12.576918
0.0150	120	0.948740	0.748954	12.327528

Таблиця 3

Results for frame duration = 0.030 (DCT-HHT)

Hop	Cutoff	Accuracy	F1_unknown	Time
0.0100	80	0.951303	0.757447	12.239028
0.0100	100	0.950021	0.753684	11.959048
0.0100	120	0.944468	0.731405	12.774837
0.0125	80	0.953439	0.764579	12.653302
0.0125	100	0.940624	0.733205	12.062211
0.0125	120	0.943614	0.715517	12.280932
0.0150	80	0.950021	0.752237	12.611945
0.0150	100	0.944468	0.724576	12.388565
0.0150	120	0.947886	0.741525	12.170956

3.3 Реалізація EEMD + Hilbert spectrum

Для порівняння реалізовано іншу, навмисно спрощену baseline, яка спирається на EEMD (Ensemble empirical mode decomposition), а потім на статистику спектру Гільберта. Детальний розв'язок також можна подивитись на github. Див. Додаток А.

3.3.1 Попередня обробка та EEMD

Аудіосигнали $x(t)$ зчитуються з частотою дискретизації $f_s = 8000$ Гц та нормалізуються. Далі застосовується **Ensemble Empirical Mode Decomposition (EEMD)**, яка декомпонує сигнал на K внутрішніх мод функцій (IMF):

$$x(t) \approx \sum_{k=1}^{K} c_k(t),$$

де $c_k(t)$ — k-та IMF, отримана за допомогою додавання білого шуму та усереднення N реалізацій (в нашому випадку $N = 20$).

3.3.2 Гільберт-перетворення і частота

Для кожної IMF $c_k(t)$ обчислюється аналітичний сигнал:

$$z_k(t) = c_k(t) + j \cdot \mathcal{H}\{c_k(t)\},$$

де $\mathcal{H}\{\cdot\}$ — перетворення Гільберта. З цього обчислюються:

- **Миттєва амплітуда**: $A_k(t) = |z_k(t)|$;
- **Миттєва фаза**: $\phi_k(t) = \arg(z_k(t))$;
- **Миттєва частота**:

$$f_k(t) = \frac{1}{2\pi} \cdot \frac{d\phi_k(t)}{dt} \approx \frac{f_s}{2\pi} \cdot \nabla\phi_k(t),$$

де f_s — частота дискретизації.

3.3.3 Статистичні ознаки спектру Гільберта

Для кожної IMF розраховуються наступні статистики:

- Середнє значення амплітуди: $\mu_A = \mathbb{E}[A_k(t)]$;
- Дисперсія амплітуди: $\sigma_A^2 = \mathrm{Var}[A_k(t)]$;

- Ентропія потужності спектру частот (за гістограмою):

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i),$$

де p_i — нормалізована потужність у i-му біні гістограми миттєвої частоти.

Ознаки для всіх IMF конкатенуються у фінальний вектор ознак $\boldsymbol{x} \in \mathbb{R}^d$.

3.3.4 Побудова датасету та масштабування

Формується набір даних:

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N},$$

де $\boldsymbol{x}_i$ — вектор ознак аудіофайлу, а $y_i \in \{0, 1\}$ — мітка класу (“unknown” або “yes_drone”).

Дані розділяються на тренувальну та тестову вибірки з подальшим масштабуванням за допомогою стандартного скейлера:

$$\boldsymbol{x}'_i = \frac{\boldsymbol{x}_i - \mu}{\sigma},$$

де μ та σ — відповідно середнє та стандартне відхилення в тренувальній вибірці.

3.3.5 Класифікація

Розглядаються три класифікатори:

1) **Підтримуючий векторний апарат (SVM)** з RBF-ядерною функцією:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|^2\right),$$

де $C = 10$, γ — “scale”.

2) **Випадковий ліс (Random Forest)** з $n = 200$ дерев та балансуванням класів.
3) **Метод k найближчих сусідів (k-NN)** з $k = 7$ та вагами, оберненими до відстані.

3.3.6 Оцінка результатів

Для кожного класифікатора розраховується точність, повнота, F-міра за кожним класом із використанням метрики `classification_report`. Загалом, цей алгоритм дає високу розрізнювальну здатність. Серед трьох протестованих моделей, випадковий ліс із 300 дерев досягає найкращої загальної точності (96.3%) та найвищого F1 для класу складних невідомих (0.834), підтримуючи при цьому повноту дронів 0,994. SVM-RBF досягає порівняної точності (95.5%), але допускає більше хибних тривог (повнота фону 0.722), тоді як k-NN (k = 7) є найшвидшою, але демонструє найслабшу фонову чутливість (повнота 0.658).

Таблиця 4

Classification Report for EEMD + Hilbert Spectrum

Model	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
SVM-RBF	unknown	0.901	0.722	0.802	266
	yes_drone	0.961	0.989	0.975	1857
	accuracy			0.955	2123
	macro avg	0.931	0.855	0.888	2123
	weighted avg	0.954	0.955	0.953	2123
RandomForest	unknown	0.943	0.748	0.834	266
	yes_drone	0.965	0.994	0.979	1857
	accuracy			0.963	2123
	macro avg	0.954	0.871	0.907	2123
	weighted avg	0.962	0.963	0.961	2123
k-NN (k=7)	unknown	0.814	0.658	0.728	266
	yes_drone	0.952	0.978	0.965	1857
	accuracy			0.938	2123
	macro avg	0.883	0.818	0.846	2123
	weighted avg	0.935	0.938	0.935	2123

РОЗДІЛ 4

АНАЛІЗ РЕЗУЛЬТАТІВ

4.1 Порівняння застосованих алгоритмів

EEMD виконує розкладання сигналу на внутрішні модальні функції (IMF), які краще виділяють нестійкі (нестаціонарні) компоненти, пов'язані з роботою ротора. Статистики з Гільберт-спектра, отримані з цих IMF, підвищують невідоме раніше значення F1-score з 0.78 (найкраще у методі DCT-HHT) до 0.834 і додають 1% абсолютної точності. Перевага досягається завдяки кращій локалізації в часі та частоті та меншій змішаності мод, що дозволяє моделі Random Forest краще розрізняти фоновий шум і гармоніки дрона.

Метод DCT-HHT, натомість, ґрунтується на одному аналітичному сигналі, сформованому після стиснення енергії в DCT-домені (дискретне косинусне перетворення). Його вектор MFCC є коротшим і обчислюється швидше, що дозволяє обробляти один файл приблизно за 0.15 секунди на мікроконтролері класу CPU — у 5–7 разів швидше, ніж спрощений варіант EEMD. Незважаючи на простішу обробку, чутливість до виявлення дронів залишається на рівні 0.99, а рівень помилкових спрацювань (FAR $\approx 4\%$) вже нижчий, ніж у багатьох систем, які поєднують радіочастотну та акустичну інформацію.

У порівнянні з методами, описаними в [18] та [19], запропонований HHT-пайплайн (на основі Гільберт-Гуанг трансформації) забезпечує подібну точність ($\approx$ 95 %), але потребує значно менше навчальних даних та менше обчислювальних ресурсів. Крім того, сучасні дослідження у сфері радіо-акустичного злиття (RF-acoustic fusion) зазвичай досягають 96–97% точності, але при цьому мають вищий рівень помилкових спрацювань (FAR); на відміну від них, поточна система зберігає FAR нижчим за 4%, що підкреслює її практичну придатність для вбудованих систем протидії БПЛА в реальному часі.

Таблиця 5

Порівняння підходів EEMD + Hilbert Spectrum та DCT + HHT (MFCC)

Критерій	**EEMD + Hilbert spectrum**	**DCT + HHT (MFCC)**
Найкраща точність	0.963 (Random Forest, 300 дерев)	0.953 (25 мс / 10 мс, 80 Гц)
F1-score (unknown)	0.834	0.780
Recall (unknown)	0.748	0.745
Recall (yes_drone)	0.994	0.990
Кількість ознак	$\leq$ 18–24 скалярів (3 статистики $\times \leq$ 6 IMF)	13 MFCC
Час добування ознак (CPU, 3 с кліп)	0.8–1.0 с (8 кГц, 20 запусків)	0.15 с
Ядро алгоритму	Адаптивне EEMD $\rightarrow$ енергозважений спектр Гільберта	DCT попередня обробка $\rightarrow$ аналітичний сигнал $\rightarrow$ MFCC
Придушення змішування мод	вбудоване (ансамбль)	не розглядається

ВИСНОВКИ

Експерименти підтверджують, що перетворення Гільберта-Хуанга є потужним засобом миттєвого захоплення характеристик нестаціонарного аудіо, що робить його добре придатним для акустичного виявлення БПЛА. Дослідження надало повне обґрунтування вибору HHT, реалізувало та протестувало алгоритм на реальному шуму дрона, а також порівняло результати з MFCC-SVM та CNN.

Хоча запропонована система вже досягає загальної точності 95% з частотою відтворення дрона 0,99, залишається кілька шляхів для вдосконалення. По-перше, невідомий клас слід збагатити та перебалансувати, доповнити фонові записи та застосувати методи передискретизації. По-друге, блок ознак можна уточнити: довжину вікна, розмір стрибка та відсічення високих частот слід налаштувати точніше; можна додати Δ- та $\Delta\Delta$-MFCC, характеристики спектрального контрасту та кольоровості; а параметри самого pipeline DCT-HHT можна скоригувати для отримання чіткіших часово-частотних структур. По-третє, класифікатор можна оновити до компактних архітектур CNN/CRNN або легких трансформерів з пакетною нормалізацією, відсіванням та ранньою зупинкою — підхід, успішно продемонстрований у мережі з низьким розміром для акустики дронів у роботі [20].

Нарешті, вихід за межі бінарного розрізнення дрон/фон може ще більше зменшити кількість хибних тривог. Багатокласова схема або стратегія виявлення аномалій можуть відокремити нетипові шумові шаблони від справжніх БПЛА; самокерована аудіовізуальна система в роботі [21] пропонує перспективний план для такого розширення.

СПИСОК ЛІТЕРАТУРИ

1. Huang, N. E. & Shen, S. S. P. *Hilbert-Huang Transform and Its Applications* ISBN: 9789812703347 (WORLD SCIENTIFIC, вер. 2005).
2. Huang, N. E. *та ін.* The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* **454,** 903—995. ISSN: 1471-2946 (бер. 1998).
3. Susanto, A. *та ін.* Application of Hilbert–Huang transform for vibration signal analysis in end-milling. *Precision Engineering* **53,** 263—277. ISSN: 0141-6359 (лип. 2018).
4. Oweis, R. J. & Abdulhay, E. W. Seizure classification in EEG signals utilizing Hilbert-Huang transform. *BioMedical Engineering OnLine* **10,** 38. ISSN: 1475-925X (2011).
5. Huang, N. E., Wu, M.-L., Qu, W., Long, S. R. & Shen, S. S. P. Applications of Hilbert–Huang transform to non-stationary financial time series analysis. *Applied Stochastic Models in Business and Industry* **19,** 245—268. ISSN: 1526-4025 (лип. 2003).
6. Bowman, D. C. & Lees, J. M. The Hilbert-Huang Transform: A High Resolution Spectral Method for Nonlinear and Nonstationary Time Series. *Seismological Research Letters* **84,** 1074—1080. ISSN: 1938-2057 (жовт. 2013).
7. Huang, N. E. & Wu, Z. A review on Hilbert-Huang transform: Method and its applications to geophysical studies. *Reviews of Geophysics* **46.** ISSN: 1944-9208 (черв. 2008).
8. Makur, A. & Mitra, S. Warped discrete-Fourier transform: Theory and applications. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* **48,** 1086—1093. ISSN: 1057-7122 (2001).
9. Çelik, A. Mathematics of Music: Chord Classification. *Kaggle.* `https://www.kaggle.com/code/ahmetcelik158/mathematics-of-music-chord-classification` (2022).

10. Ko, B. J. *та ін. Acoustic Signal Processing for Anomaly Detection in Machine Room Environments: Demo Abstract* в *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments* (ACM, листоп. 2016), 213—214.

11. Mezei, J., Fiaska, V. & Molnar, A. *Drone sound detection* в *2015 16th IEEE International Symposium on Computational Intelligence and Informatics (CINTI)* (IEEE, листоп. 2015).

12. Alam, M. J., Kenny, P., Dumouchel, P. & O'Shaughnessy, D. *Robust speech recognition using warped DFT-based cepstral features in clean and multistyle training* в (IEEE, Lisbon, Portugal, 2014), 1791—1795. ISBN: 978-0-9928-6261-9.

13. Kulkarni, P., Kulkarni, S., Mulange, S., Dand, A. & Cheeran, A. N. *Speech recognition using wavelet packets, Neural Networks and Support Vector Machines* в *2014 International Conference on Signal Propagation and Computer Technology (ICSPCT 2014)* (IEEE, лип. 2014), 451—455.

14. Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America* **87,** 1738—1752. ISSN: 1520-8524 (квіт. 1990).

15. Stevens, S. S. On the psychophysical law. *Psychological Review* **64,** 153—181. ISSN: 0033-295X (1957).

16. Chaudhari, H., Nalbalwar, S. L. & Sheth, R. *A review on intrensic mode function of EMD* в *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (2016), 2349—2352.

17. Al-Emadi, S. A., Al-Ali, A. K., Al-Ali, A. & Mohamed, A. Audio Based Drone Detection and Identification using Deep Learning. *IWCMC 2019 Vehicular Symposium (IWCMC-VehicularCom 2019).* `https://github.com/saraalemadi/DroneAudioDataset/tree/master` (2019).

18. Mrabet, M., Sliti, M. & Ammar, L. B. Machine learning algorithms applied for drone detection and classification: benefits and challenges. *Frontiers in Communications and Networks* **5.** ISSN: 2673-530X (жовт. 2024).

19. Kim, J., Kim, Y., Shin, H., Wang, Y. & Matson, E. *How Far Can a Drone be Detected? A Drone-to-Drone Detection System Using Sensor Fusion* в *Proceedings of the 15th International Conference on Agents and Artificial Intelligence* (SCITEPRESS - Science та Technology Publications, 2023).

20. Aydın, İ. & Kızılay, E. Development of a new Light-Weight Convolutional Neural Network for acoustic-based amateur drone detection. *Applied Acoustics* **193,** 108773. ISSN: 0003-682X (трав. 2022).

21. Xiao, Z., Yang, Y., Xu, G., Zeng, X. & Yuan, S. *AV-DTEC: Self-Supervised Audio-Visual Fusion for Drone Trajectory Estimation and Classification* 2024.

ДОДАТОК А

Посилання на код у GitHub:

https://github.com/marmurr/HHT_drone_detection_on_audio