

АКУСТИЧНА КЛАСИФІКАЦІЯ БПЛА В УМОВАХ ДОМЕННОГО ЗСУВУ ТА ДЕФІЦИТУ ДАНИХ

Вільгурін В.І.¹, Сидорський В.С.², Данилов В.Я.³

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

¹ vvilgurin@gmail.com, ² volodymyr.syderskyi@gmail.com, ³ danilov1950@ukr.net

Зростання кількості та доступності безпілотних літальних апаратів (БПЛА) створює нові виклики для безпеки цивільної та критичної інфраструктури. Одним з перспективних напрямів є акустичне виявлення БПЛА, оскільки мікрофони є дешевими, пасивними сенсорами і можуть розгортатися на великій території. Водночас практичне застосування таких систем часто ускладнюється відмінностями у фізичних властивостях мікрофонів, різноманіттю умов навколишнього середовища та обмеженим обсягом анотованих цільових даних. У роботі розглянуто підхід до побудови моделі акустичної класифікації БПЛА, стійкої до доменного зсуву, що виникає внаслідок відмінностей у мікрофонах та локаціях і дефіциту маркованих прикладів. Запропонована система поєднує попередню обробку сигналу на основі мел-спектрограм, використання попередньо натренованої згорткової мережі як екстрактора ознак, агрегацію на основі механізму уваги та стратегії адаптації домену із застосуванням аугментацій та збалансованого відбору даних.

Ключові слова: акустична детекція, БПЛА, ConvNeXt, доменний зсув, Mixup, SED.

1. ВСТУП

Акустичне виявлення безпілотних літальних апаратів є важливим компонентом сучасних систем спостереження та безпеки. На відміну від радіолокаційних чи оптичних підходів, акустичні системи не потребують прямої видимості цілі та можуть працювати в умовах складного рельєфу або задимлення, а також не можуть бути легко виявлені в силу пасивної природи роботи.

Водночас вони є чутливими до типу мікрофона, акустики сцени та рівня фонового шуму. На практиці це призводить до доменного зсуву: дані, зібрані на різних пристроях або в інших умовах, мають інший розподіл, ніж той, на якому навчалася модель, а також, потенційно, інший набір класів. Крім того, зібрати велику кількість якісних маркованих записів БПЛА в реалістичних умовах часто неможливо через високу складність анотації в офлайн-режимі та високу вартість експериментів.

Таким чином система повинна ефективно використовувати інформацію з наявних даних для адаптації до нових доменів та нових класів. Метою роботи є розробка акустичної системи класифікації БПЛА, яка:

- зберігає високу точність у присутності доменного зсуву між мікрофонами та акустичними сценами;
- враховує сильний дисбаланс класів (більшість записів-природний шум, меншість-БПЛА);

- залишається обчислювально ефективною для застосування в режимі, близькому до реального часу.

2. ПОСТАНОВКА ЗАДАЧІ ТА ДАНІ

Розглядається задача класифікації аудіозаписів на основі коротких фрагментів сигналу. Вхідними даними є хвильові форми, записані щонайменше двома типами мікрофонів, що відрізняються частотною характеристикою, рівнем власного шуму та акустичними умовами. Розподіл даних наведено в таблиці 1.

Таблиця 1. Розподіл даних за доменами та класами

Домен (Набір даних)	Шум	Не БПЛА	БПЛА	Усього
Мікрофон 1 (Train)	299 603	17 981	0	317 584
Мікрофон 2 (Train)	2 519	0	441	2 960
Мікрофон 1 (Val)	43 551	2 855	0	46 406
Мікрофон 2 (Val)	2 797	0	4 207	7 004
Тестовий набір	343	0	805	1 148

Спостерігається значний дисбаланс між кількістю наявних аудіозаписів на рівні мікрофонів та на рівні класів, різні акустичні сцени та обмежений обсяг маркованих прикладів у цільовому домені (рис. 1).

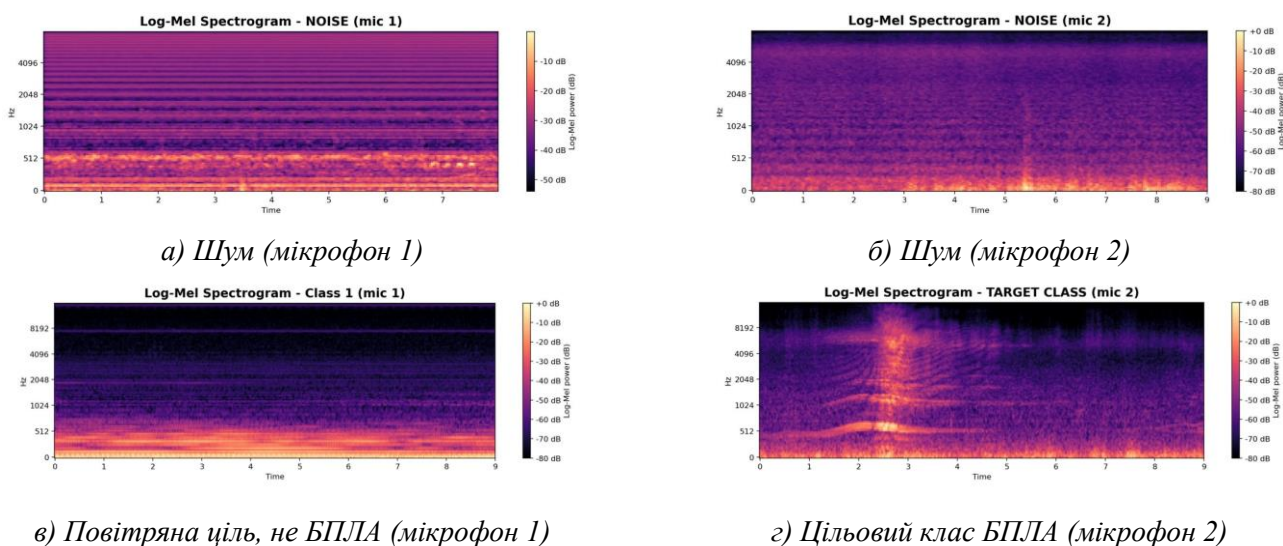


Рисунок 1. Порівняння спектральних характеристик сигналів

Необхідно побудувати модель f , яка на основі часово-частотного представлення аудіосигналу прогнозує ймовірність належності до класу БПЛА, залишаючись стійкою до зміни домену та використовуючи обмежену кількість маркованих записів з нового середовища.

3. МЕТОДИ ТА АРХІТЕКТУРА МОДЕЛІ

Попередня обробка. Для побудови ознак використано перетворення звукових хвиль з частотою дискретизації 32 кГц у мел-спектрограми. Аудіо ділиться на фрагменти тривалістю 9 с, до яких застосовується короткочасне перетворення Фур'є (STFT) з параметрами $n_{fft} = 2048$ та $hop = 512$, після чого спектр перетворюється у мел-шкалу [1]. Така попередня обробка забезпечує достатню частотну роздільну здатність для гармонік цільових сигналів.

Модель. У ролі енкодера використано згорткову мережу ConvNeXt-Tiny [2], попередньо натреновану на великому наборі даних з зображеннями. Верхні шари ConvNeXt виступають

екстрактором спектрально-часових ознак. Сформовані ознаки подаються до спеціалізованої голови типу SED (Sound Event Detection) [3], реалізованої як уваговий модуль. На першому етапі відбувається агрегація по частоті за допомогою GeM-pooling, яке дозволяє більш гнучко підлаштовуватись до різного спектрального розподілу сигналів з двох доменів. Далі часові вектори обробляються рекурентним шаром GRU, що використовується для моделювання темпоральних змін у спектрограмі. Врешті, отримані ознаки зважуються механізмом уваги по часовій осі для визначення важливості кожної з частин зображення для покращеного узагальнення у разі виникнення високоамплітудного шуму.

Доменна адаптація. Для підвищення стійкості до доменного зсуву застосовано комбінацію аугментацій на рівні хвильової форми та спектрограм. На рівні сигналу виконуються нормалізація амплітуди, випадкове інвертування сигналу в часі, *gandom gain* та *time-stretch*. Також виконується змішування шумових записів з усіма іншими прикладами (Mixup [4]):

$$x_{mix} = \lambda x_i + (1 - \lambda) x_j, \quad y_{mix} = y_i \quad (1)$$

де $\lambda \sim \text{Beta}(\alpha, \alpha)$. При цьому для зразків з мікрофона цільового домену не застосовуються перетворення, які суттєво змінюють частотну характеристику спектрограми, через брак даних цільового домену. Маркування завжди залишається «жорстким»: усі змішані приклади вважаються тим самим класом, що й основний сигнал, за винятком випадку, коли шум другого мікрофона змішується з шумом першого – тоді зразок позначається як шум.

На рівні спектрограм застосовуються маскування часу та частоти (SpecAugment), що визначається як добуток Адамара спектрограми $S(f, t)$ та бінарних масок M :

$$\tilde{S}(f, t) = S(f, t) \odot M_f(f) \odot M_t(t), \quad (2)$$

а також невеликий випадковий зсув, що додатково зменшує перенавчання на конкретні артефакти домену та випадковий шум.

Тренування. Навчання виконується з використанням категорійної крос-ентропійної функції втрат. Для боротьби з дисбалансом класів застосовується зважений семплінг: ймовірність вибору зразків різних класів масштабується через квадратний корінь від їхньої частоти, що пом'якшує домінування шуму та дозволяє моделі краще навчатися на рідкісних корисних класах і цільовому класі БПЛА. Така комбінація фронтенду на мел-спектрограмах, ConvNeXt бекбону, увагової SED-голови та спеціалізованих аугментацій робить модель придатною для роботи в умовах доменного зсуву та дефіциту маркованих даних цільового домену.

4. ЕКСПЕРИМЕНТИ ТА РЕЗУЛЬТАТИ

Для оцінки ефективності запропонованого підходу було проведено серію експериментів. Було здійснено оцінку впливу різної кількості класів для навчання та методів аугментації на здатність моделі до узагальнення на тестовій вибірці (таблиця 2).

Таблиця 2. Результати експериментів на тестовій вибірці

Експеримент	Precision	Recall	F1 Score
Baseline (Всі дані)	0.989	0.908	0.947
Всі дані + Mixup	0.954	0.958	0.956
Тільки Мікрофон 2	0.989	0.810	0.891
Pre-train + Fine-tune	0.995	0.903	0.947

1. Навчання лише на цільовому домені (Мікрофон 2). У першому експерименті модель навчалася лише на обмеженому наборі даних з другого мікрофона із застосуванням базових аугментацій. Результати виявилися незадовільними (Recall 0.810) порівняно з іншими підходами. Аналіз показує, що модель схильна до перенавчання, орієнтуючись на прості ознаки, такі як загальний баланс високих та низьких частот, замість вивчення специфічних гармонійних патернів БПЛА.

2. Об'єднання доменів (Мікрофон 1 + Мікрофон 2). Додавання великого масиву шумових даних з першого мікрофона покращило метрики якості. Проте аналіз карт активації (GradCAM) продемонстрував, що модель все ще часто фокусується на низькочастотних областях як для класу шуму, так і для класу сигналу, що знижує надійність детекції в умовах сильного зашумлення.

3. Об'єднання доменів із застосуванням Міхур. Найкращі результати (F1 0.956) були отримані при використанні техніки Міхур, де шуми з другого мікрофона змішувалися з цільовими сигналами та сигналами з мікрофона 1. Модель вчилася шукати корисні ознаки не тільки серед низькочастотного шуму, а й у середньому та високому частотних діапазонах та виокремлювати ознаки, специфічні для кожного з класів.

4. Попереднє навчання з донавчанням на цільовому домені (Pre-training+Fine-tuning). В рамках цього експерименту було застосовано стратегію трансферного навчання. На етапі Pretraining модель навчалася на повному об'єднаному наборі даних з Міхур, що дозволило вивчити робастні ознаки. На етапі Fine-tuning модель донавчалася виключно на даних цільового домену. Цей підхід забезпечив високу точність та адаптацію до умов цільового мікрофона. Однак, донавчання на малій вибірці призвело до того, що модель стала більш впевненою на простих прикладах, але надмірно консервативною на складних або зашумлених зразках.

Аналіз передбачень. Детальний аналіз помилок моделі дозволив виявити критичну проблему розбіжності між тренувальним та тестовим наборами даних (рис. 2). Тренувальна вибірка складається переважно із записів БПЛА на близькій відстані до мікрофону, які характеризуються широким спектральним складом та чіткими високочастотними гармоніками. Натомість тестовий набір містить записи віддалених дронів. Через фізичні властивості, у сигналах віддалених об'єктів високочастотні компоненти швидко загасають, залишаючи переважно низькочастотні артефакти. Оскільки в процесі навчання модель бачила низькочастотну активність переважно у класі “шум”, а сигнал БПЛА асоціювала з наявністю вищих частот, вона не здатна ефективно відрізнити віддалений дрон від фонового шуму. Це свідчить про необхідність розширення вибірки реальними записами віддалених цілей для покращення результатів.

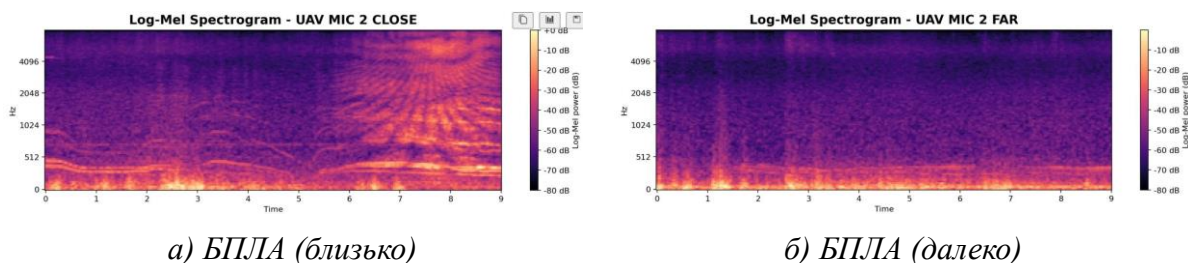


Рисунок 2. Порівняння спектрограм БПЛА у залежності від відстані до мікрофона

5. ВИСНОВКИ

У роботі досліджено проблему акустичної класифікації БПЛА в умовах обмежених даних цільового домену та використання різнотипного обладнання. Експерименти показали, що

навчання виключно на малій вибірці цільового мікрофона призводить до перенавчання на поверхневі ознаки та низької здатності до узагальнення.

Найкращу ефективність (F1-score = 0.956) продемонстрував підхід із глобальним навчанням на об'єднаних даних та використанням Міхур. Ця стратегія змушує модель ігнорувати низькочастотний шум і фокусуватися на стійких спектральних гармоніках БПЛА. Критичним викликом залишається фізична відмінність між тренувальними (ближніми) та тестовими (віддаленими) сигналами, де втрата високих частот ускладнює детекцію. Подальша робота буде зосереджена на вдосконаленні методів попередньої обробки для покращення детекції віддалених цілей.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Qiuqiang Kong та ін. “PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition”. В: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28(2020), с. 2880–2894.
2. Zhuang Liu та ін. “A ConvNet for the 2020s”. В: *arXiv preprint arXiv:2201.03545* (2022).
3. Krishna Kumar Singh та Yong Jae Lee. “Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization”. В: *arXiv preprint arXiv:1710.02998* (2017).
4. Hongyi Zhang та ін. “mixup: Beyond Empirical Risk Minimization”. В: *arXiv preprint arXiv:1710.09412* (2017).