https://www.sworldjournal.com/index.php/swj/article/view/swj35-01-083
DOI: 10.30888/2663-5712.2026-35-01-083

УДК 004.896

COMPARATIVE ANALYSIS OF MODERN COMPONENTS OF INFORMATION TECHNOLOGY FOR SOUND PROCESSING ON ROBOTIC SYSTEMS

ПОРІВНЯЛЬНИЙ АНАЛІЗ СУЧАСНИХ КОМПОНЕНТІВ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ОБРОБКИ ЗВУКУ НА РОБОТИЗОВАНИХ СИСТЕМАХ

Korotenko G.M. / Коротенко Г.М.
d.t.s., prof. / д.т.н., проф.
ORCID: 0000-0003-3774-5260
Fridman K.V. / Фрідман К.В.
student / студент
Dnipro University of Technology, D. Yavornitsky, avenue, 19, Dnipro 49600, Ukraine
Национальный технический университет «Днепровская политехника», пр-т Дмитрия Яворницкого, 19, 49600, г. Днепр, Украина

Аннотация. *В роботі розглядається сучасний стан і можливості обробки інформаційних потоків у сучасних роботизованих системах із засобами штучного інтелекту. Наведено порівняльні характеристики компонентів різних систем і технологій для оцінки можливостей їхнього використання у різних умовах.*

Ключевые слова: *штучний інтелект, роботизовані системи, звукова інформація, обробка інформації, хмарний сервіс, математичні моделі.*

Вступ.

Останнім часом у світі штучного інтелекту (ШІ) та роботизованих систем відбувається революція, пов’язана зі стрімким зростанням можливостей вбудованих систем та появою просунутих математичних моделей, які здатні не просто прогнозувати певні значення на підставі комбінацій вхідних даних, але й розуміти контекст цих даних, складні взаємозв’язки тощо. Сучасні роботи вже змагаються за право першості у плавності рухів та розумінні того, що відбувається навколо них. І таке розуміння неможливе без повноти збору інформації про навколишнє середовище, де, серед іншого, важливою може бути звукова інформація. Її збір та обробка може вчасно повідомити модуль прийняття рішень роботизованої системи про можливу загрозу або необхідність вчасно відреагувати.

Попри те, що апаратні можливості роботизованих систем невпинно збільшуються, складність сучасних систем штучного інтелекту збільшується набагато швидше. Для того щоб мати можливість враховувати результати

обробки будь-якого різновиду інформації, який породжує великі дані, наприклад, відеопотоку з камер спостереження, аудіопотоку з мікрофонів, геолокаційних та інших даних з вбудованих датчиків, робот чи безпілотна система має дві можливості: або обробити дані на борту, або відправити їх на обробку на хмару. Хмарна обробка широко вживається у персональних та бізнес-застосунках в умовах стабільного каналу комунікацій та за відсутності суттєвих обмежень у часі. Однак якщо роботизована або безпілотна система опиняється в умовах обмеження часу або можливості передати дані на обробку, такий підхід може не спрацювати, й тоді постає питання: чи здатна ця роботизована або безпілотна система забезпечити транспортування достатньо потужного комп'ютера та джерела живлення, аби мати можливість виконати обробку на борту.

Таким чином, актуальною є задача визначення архітектури інформаційної технології обробки звуку, яка може бути використана на роботизованих системах. Ця обробка може відбуватися в умовах без обмежень ресурсів часу або з такими обмеженнями. Також додатково слід врахувати можливість впливу засобів радіоелектронної боротьби та інших завад природного та техногенного характеру. І першим кроком для визначення / обрання архітектури інформаційної технології є проведення порівняльного аналізу елементів, що потенційно можуть увійти до складу такої технології, задля визначення перспективної конфігурації архітектури, чому й присвячена ця публікація.

Основна частина.

Для виконання порівняльного аналізу стану розвитку потенційних компонентів можливої інформаційної технології обробки звуку на роботизованих системах дослідімо характеристики підсистем комунікацій та математичних моделей обробки звуку, що часто використовуються у подібних технологіях. Дані, що будуть наведені, взяті із відкритих джерел та наукових публікацій, присвячених математичним моделям обробки звуку. Результати розгляду стану розвитку цих компонентів має надати можливість оцінки, чи можуть роботи обробляти звук самостійно або вони мають тільки збирати й

передавати інформацію на хмарний сервіс. Також може бути поставлене питання про комбіновану архітектуру, де використання попередньо навчених моделей відбувається на роботі з одночасним донавчанням моделей на хмарному сервісі.

Якщо оцінювати можливість використання віддаленого хмарного сервісу для обробки звуку, зазвичай варто виконати оцінку пропускної спроможності каналу зв'язку, реальної латентності цього каналу, а також усієї системи загалом, що включає час на локальну обробку на роботі, та час реакції системи штучного інтелекту на хмарному сервісі. Враховуючи характер різних задач, пропускна спроможність каналу зв'язку повинна відповідати певним вимогам [1], які наведені у таблиці 1.

Таблиця 1 - Трафікові вимоги для передачі аудіо різної якості

Аудіо-параметри	Частота дискре-тизації	Дискрет-ність	Канали	Бітрейт (нестис-нений)	Бітрейт (стиснене)	Практичне застосування
Якість телефонного зв'язку	8 кГц	8 біт	Моно	64 кбіт/с	16-32 кбіт/с	Базові голосові команди
Якість розпізнавання голосу	16 кГц	16 бит	Моно	256 кбит/с	32-64 кбит/с	Голосові помічники
CD якість	44.1 кГц	16 бит	Стерео	1,411 кбит/с	128-256 кбит/с	Музика
Професійна якість	48 кГц	16 бит	Моно	768 кбит/с	96-192 кбит/с	Класифікація звуків середовища
Високоякісна класифікація	48 кГц	24 бит	Стерео	2,304 кбит/с	256-384 кбит/с	Детальна класифікація загроз

Авторська розробка

З наведених у таблиці 1 даних можна зробити висновок, що впевнена ідентифікація загроз може вимагати доволі широкого каналу зв'язку, тим більше якщо для розуміння контексту не можна пригнічувати сторонні шуми. Латентність каналів зв'язку, з різними характеристиками, які найчастіше використовуються роботизованими системами наведені у таблиці 2.

Таблиця 2 - Характеристики типів каналів зв'язку для роботизованих систем

Тип каналу зв'язку	Дальність	Частотний діапазон	Типова латентність	Пропускна спроможність
Прямий радіоканал (2,4 ГГц)	1-5 км	2.4-2.5 ГГц	50-100 мс	1-10 Мбит/с
Прямий радіоканал (5,8 ГГц)	0.5-3 км	5.725-5.875 ГГц	30-80 мс	10-50 Мбит/с
Мобільний зв'язок 4G/LTE	Залежить від покриття	700 МГц - 2.6 ГГц	30-70 мс	5-50 Мбит/с
Мобільний зв'язок 5G	Залежить від покриття	Sub-6 ГГц, mmWave	10-30 мс	50-1000 Мбит/с
Супутники Starlink	Глобальна	Ku-band (10,7–14,5 ГГц), Ka-band (17,8–30 ГГц)	20–60 мс	50–280 Мбіт/с
Супутники LEO	Глобальна	L-band(1–2 ГГц), Ka-band (26,5–40 ГГц)	20-50 мс	50-500 Мбит/с
Супутники GEO	Глобальна	C-band(4–8 ГГц), Ku-band(12–18 ГГц), Ka-band(26,5–40 ГГц)	500-700 мс	1-100 Мбит/с

Авторська розробка

Дані в таблиці наведено без урахування перешкод викликаних використанням засобів радіоелектронної боротьби, або інших джерел завад техногенного характеру. Для якісної ідентифікації загроз бажано забезпечити канал зв’язку від 256 кбіт/с або вище. Такій вимозі відповідають всі наведені типи каналів зв’язку. І в такому разі основними факторами добору є пряма радіовидимість, латентність та можливість працювати в умовах радіоелектронних завад. З наведених типів каналів зв’язку тільки супутникові технології є відносно не чутливими до засобів радіоелектронної боротьби, всі решта технологій можуть бути доволі легко придушені, тож для задач цивільного характеру треба враховувати виключну архітектуру самої роботизованої системи, тобто чи є пряма радіовидимість і чи достатньо латентності обраного прямого каналу для виконання поставлених задач. Або у більш універсальному

випадку можна обирати канали 4G чи 5G в тих районах світу, де такі технології вже розгорнуті. Для задач воєнного характеру перевагу слід віддати супутниковим технологіям, проте вибір конкретної з них слід провести після аналізу поставленої задачі та технічних умов її виконання.

Латентність на пристрої, як правило визначається тим, чи використовується компресія та пригнічення шуму, чи ні. В будь якому разі, сучасні мобільні процесори мають достатньо високу продуктивність для швидкого розв'язання таких задач. Крім того, можуть використовуватися додаткові цифрові процесори обробки сигналів (DSP), які зазвичай беруть на себе спеціальні задачі з обробки звуку, що підвищує загальну продуктивність системи. Латентність сервера визначається його режимом експлуатації (виділений або спільний) та складністю математичних моделей для розв'язку задач. Як правило, на серверах загального призначення немає спеціальних пристроїв, які здатні виконувати попередню обробку звуку, тому подібні задачі, як правило, розв'язуються шляхом використання додаткових програмних засобів, що може загалом негативно впливати на латентність системи.

Математичні моделі для обробки звуку на роботизованій системі мають відповідати певним критеріям для того аби вони могли бути запущені на відповідному мобільному пристрої та їхня робота могла бути ефективною. Серед визначальних характеристик для цих моделей є можливість роботи на процесорах чи мікроконтролерах, що найчастіше використовуються під час побудови мобільних систем штучного інтелекту, відносно невеликий розмір у пам'яті, низька латентність за умови прийнятної точності тощо. Основні моделі, що відповідають цим критеріям, наведено у таблиці 3.

У якості джерела даних, що використовувалися під час навчання та тестування моделей з таблиці 3, використовуються набори даних, характеристики яких наведено у таблиці 4.

Розгляньмо коротко характеристики моделей з таблиці 3.

Micro-ACDNet [2] - ультракомпактна модель, розроблена спеціально для мікроконтролерів класу ARM Cortex. Має лише 131 тисячу параметрів та займає

0,5 МБ пам'яті, що є найменшим показником серед розглянутих моделей. Ключовою особливістю є робота безпосередньо з raw audio сигналом частотою 20 кГц, що усуває потребу в обчислювально затратному перетворенні у спектрограму на пристрої. Модель досягає точності 83,65% на наборі даних ESC-50, 96.25% на ESC-10, а також 78.28% на UrbanSound8K та є оптимальним вибором для систем із жорсткими обмеженнями енергоспоживання.

Таблиця 3. - Математичні моделі обробки звуку що найчастіше використовуються на роботизованих системах

Модель	Пара-метри	Розмір	FLOPs	ESC-	Латент-ність	Плат-форма	Область застосування
Micro-ACDNet	0.131M	0.5 MB	14.82M	ESC-50: 83.65%	<10 ms	ARM Cortex MCU	Виявлення звукових подій в реальному часі на MCU; ультра-низьке енергоспоживання
YAMNet-256	~0.3M	~1 MB	~50M	ESC-10: 96%	<15 ms	STM32 MCU	Класифікація environmental sounds; IoT пристрої; смарт-датчики
MN-04 (Efficient AT)	0.983M	~4 MB	110M	N/A	~20 ms	Mobile GPU/NPU	Аудіо-тегування на мобільних пристроях; real-time моніторинг
DyMN-S (DyMN-04)	1.97M	~8 MB	120M	ESC-50: 96.4%	~25 ms	Mobile GPU/NPU	Динамічна класифікація з адаптивною складністю; енергоефективність
ACDNet	4.74M	18 MB	544M	ESC-50: 87.10%	~50 ms	Edge GPU (Jetson)	Висока точність на edge; автономні роботи; дрони з GPU
ESC-NAS	~0.5M	<2 MB	~30M	ESC-50: 81%	<15 ms	MCU (NAS-opt)	Оптимізовано для конкретного MCU; hardware-aware deployment

Авторська розробка

YAMNet-256 [3, 4] - оптимізована версія моделі YAMNet від Google, адаптована компанією STMicroelectronics для розгортання на мікроконтролерах серії STM32. Має близько 300 тисяч параметрів та використовує mel-спектрограми розміром 96×64. Модель попередньо навчена на AudioSet та демонструє точність 88.75% на ESC-50, а також 94.9% на ESC-10. Перевагою є

широка підтримка інструментів STM32Cube.AI для автоматичної оптимізації та розгортання на конкретних MCU.

MN-04 (EfficientAT) [5, 6] - компактна модель із серії EfficientAT, що базується на архітектурі MobileNetV3 з адаптацією для аудіо. Містить 983 тисячі параметрів та досягає 43.2% на AudioSet, що є одним із найкращих показників серед моделей такого розміру, а також 93.2% на ESC-50. Використовує mel-спектрограми зі 128 bins та hop size 10 мс. Призначена для мобільних GPU та NPU, забезпечуючи баланс між точністю та швидкодією.

DyMN-S (DyMN-04) [5, 6] - модель із динамічною архітектурою, що адаптує обчислювальну складність залежно від складності вхідного сигналу. Має 1,97 мільйона параметрів та досягає 45% на AudioSet та 96.4% на ESC-50. Ключовою інновацією є механізм dynamic inference, який дозволяє зменшувати енергоспоживання на простих семплах та задіювати повну потужність моделі лише для складних випадків. Це особливо корисно для роботизованих систем з обмеженим бюджетом енергії.

ACDNet [2] - повнорозмірна версія архітектури Attention-based Convolutional Dilated Network із 4,74 мільйона параметрів та розміром 18 МБ. Демонструє найвищу точність серед розглянутих edge-моделей: 87,1% на ESC-50 та 96.65% на ESC-10. Використовує техніку «dilated convolutions» для збільшення рецептивного поля без зростання кількості параметрів та механізм «attention» для фокусування на релевантних частинах спектрограми. Потребує «edge GPU» типу NVIDIA Jetson для ефективної роботи.

Таблиця 4 - Порівняльна таблиця наборів даних для навчання систем ШІ з обробки звуку.

Набір даних	Кількість кліпів	Кількість класів	Тривалість	Тип міток	Метрика
AudioSet	2M+	527	10 сек	multi-label, weak	mAP
ESC-50	2,000	50	5 сек	single-label	Accuracy
ESC-10	400	10	5 сек	single-label	Accuracy

Авторська розробка

ESC-NAS [7] - Оптимізована для конкретних MCU платформ з урахуванням реальних обмежень пам'яті та обчислювальних ресурсів. Має близько 500 тисяч параметрів та досягає точності 81% на ESC-50, 96,25% на ESC-10. Перевагою є те, що архітектура враховує не лише точність, але й реальну швидкодію та енергоспоживання на цільовій платформі.

З наведених характеристик та опису випливає, що на вибір певної моделі, що буде використовуватися на деякому пристрої, впливає розмір моделі, а також можлива спеціалізація під архітектурні особливості мікропроцесорного модуля. Також з усіх наведених моделей, лише модель ACDNet вимагає наявності повноцінного процесора штучного інтелекту на кшталт Jetson Nano від Nvidia, або подібних пристроїв від інших виробників.

Добір потенційних математичних моделей для використання на хмарному сервісі був здійснений в переліку наведеному в таблиці 5.

Таблиця 5 -Математичні моделі обробки звуку що найчастіше використовуються на хмарних сервісах.

Модель	Параметри	Точність AudioSet	Точність ESC-50
BEATs (iter3+)	90M	48.6%	98.1%
EAT-base	88M	48.6%	95.9%
AST	86M	45.9%	95.6%
HTS-AT	31M	47.1%	97.0%
Audio-MAE	86M	47.3%	94.1%
PANN CNN14	81M	43.1%	94.7%

Авторська розробка

Відзначимо, що в таблиці 5 дані стосовно точності наведені тільки для наборів даних ESC-50 та AudioSet. Для відносно великих моделей набір даних ESC-10 майже не використовується для визначення точності. Ймовірно це пояснюється невеликою кількістю класів.

Розгляньмо коротко характеристики моделей з таблиці 5.

BEATs (iter3+) [8] - модель для класифікації звуку, розроблена дослідницьким підрозділом Microsoft Research. Містить 90 мільйонів параметрів

та є найточнішою серед розглянутих audio-only архітектур. Ключовою особливістю є ітеративний процес самонавчання, де модель та акустичний токенізатор почергово покращують один одного протягом кількох циклів. Досягає точності 48.6% на AudioSet та 98,1% на ESC-50. Модель є оптимальним вибором для систем із максимальними вимогами до точності класифікації.

EAT-base [9] - ефективний аудіотрансформер, розроблений Shanghai Jiao Tong University для швидкого попереднього навчання. Має 88 мільйонів параметрів та базується на архітектурі ViT-B. Ключовою перевагою є швидкість навчання, яка у 15 разів вища порівняно з BEATs завдяки оптимізованій стратегії маскування патчів. Досягає точності 48.6% на AudioSet та 95,9% на ESC-50. Модель є оптимальним вибором для сценаріїв, де потрібне швидке донавчання на нових даних.

AST (Audio Spectrogram Transformer) [8] - перша повністю attention-based модель для класифікації аудіо, розроблена в MIT. Містить 86 мільйонів параметрів та не використовує згорткових шарів. Ключовою особливістю є адаптація компонента Vision Transformer шляхом розбиття спектрограми на патчі розміром 16×16 та ініціалізація вагами з ImageNet. Досягає точності 45.9% на AudioSet та 95,6% на ESC-50. Модель має широку підтримку в екосистемі HuggingFace, що спрощує інтеграцію в існуючі системи.

HTS-AT [8] (Hierarchical Token-Semantic Audio Transformer) - компактна модель із ієрархічною структурою, розроблена для класифікації та локалізації звукових подій. Має лише 31 мільйон параметрів, що є найменшим показником серед розглянутих серверних моделей. Ключовою особливістю є використання віконної уваги замість глобальної, що зменшує обчислювальну складність та потребу в GPU-пам'яті утричі порівняно з AST. Досягає точності 47.1% на AudioSet та 97,0% на ESC-50. Модель є оптимальним вибором для серверів із обмеженими ресурсами та задач виявлення звукових подій у часі.

Audio-MAE [8] - модель на основі маскованого автокодувальника, адаптованого для аудіо домену компанією Meta AI. Містить 86 мільйонів параметрів та використовує архітектуру енкодер-декодер. Ключовою

особливістю є маскування 80% патчів спектрограми під час навчання, що змушує модель вивчати семантично значущі репрезентації без розмічених даних. Досягає точності 47.3% на AudioSet та 94,1% на ESC-50. Модель є оптимальним вибором для доменів із обмеженою кількістю розмічених аудіозаписів.

PANN CNN14 [10] - попередньо навчена згорткова нейронна мережа з 14 шарами, розроблена в Університеті Саррея. Містить 81 мільйон параметрів та базується виключно на CNN-архітектурі без механізмів уваги. Ключовою особливістю є стабільна та передбачувана робота, що зробило модель стандартним baseline для порівняння нових архітектур. Досягає точності 43.1% на AudioSet та близько 94.7% на ESC-50. Модель є оптимальним вибором для швидкого прототипування та використання як екстрактор ознак для інших задач.

З огляду на короткий опис наведених моделей можна зробити висновок, що за своїми характеристиками точності наведені моделі є зіставними та вибір конкретної моделі повинен відбуватися з огляду на характер вхідних даних та апаратні ресурси сервера де буде розгорнуто відповідну модель.

Підсумок та висновки.

З наведених вище порівнянь компонентів системи обробки звуку на роботизованих системах, можна зробити висновок, що порівняно невеликі за розміром платформи здатні забезпечити невеликий енергетичний бюджет, майже напевно не вдасться повноцінно використовувати для самостійних операцій з обробки звуку, за винятком ситуацій коли кількість класів розпізнавання буде мала, а середовище буде відносно чистим від сторонніх шумів. В цьому випадку такі платформи можуть бути споряджені енергоощадними чіпами на кшталт STM32, які здатні будуть виконати ці обмежені операції з класифікації звукових подій. Використання хмарного сервісу разом із такими платформами доцільно лише в тому випадку, якщо він виконує весь комплекс дій з обробки звуку. Фактично така роботизована система буде виконувати функцію мобільного мікрофона. В такому випадку, очевидно, що подібна система буде дуже чутлива до якості каналу зв’язку. Вона може використовуватися або у якості допоміжної підсистеми, яка немає суттєвого

впливу на характер поведінки роботизованої системи загального призначення, або винятково на цивільних роботизованих системах, які працюють в межах стабільного покриття бездротових каналів зв'язку.

Середні та великі за розміром роботизовані системи, що здатні забезпечити живлення платформ зі встановленими процесорами штучного інтелекту, здатні забезпечити автономні процеси обробки звуку з використанням заздалегідь навчених моделей. При цьому хмарний сервіс може бути використаний для донавчання наявної моделі, або навчання нової моделі задля усунення ефекту накопичувальної помилки.

З огляду на викладені вище міркування, перспективними з точки зору побудови архітектури інформаційної технології, виглядає спільне використання моделі ACDNet разом з HTS-AT для побудови розподіленої системи обробки звуку загального призначення на роботизованих системах середнього розміру. Однак, сполучення цих моделей, або будь-яке інше сполучення, в наведених у цій статті моделей, потребує додаткових експериментів у майбутньому.

Література:

1. Steven W. Smith. The Scientist and Engineer's Guide to Digital Signal Processing / Audio Processing. Chapter 22. - P. 353-372. URL: https://www.analog.com/media/en/technical-documentation/dsp-book/dsp_book_Ch22.pdf

2. Environmental Sound Classification on the Edge: A Pipeline for Deep Acoustic Networks on Extremely Resource-Constrained Devices. arXiv preprint. 2021. URL: https://arxiv.org/abs/2103.03483

3. TensorFlow. Transfer Learning for Audio Recognition. URL: https://www.tensorflow.org/tutorials/audio/transfer_learning_audio

4. STMicroelectronics. STM32 AI Model Zoo. URL: https://github.com/STMicroelectronics/stm32ai-modelzoo

5. EfficientAT: GitHub Repository. URL: https://github.com/fschmid56/EfficientAT

6. Efficient Large-Scale Audio Tagging via Transformer-to-CNN Knowledge Distillation. arXiv preprint. 2023. URL: https://arxiv.org/abs/2310.15648

7. ESC-NAS: Environment Sound Classification Using Hardware-Aware Neural Architecture Search for the Edge. Sensors. 2024. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11207705/

8. BEATs: Audio Pre-Training with Acoustic Tokenizers. arXiv preprint. 2022. URL: https://arxiv.org/abs/2212.09058

9. EAT: Self-Supervised Pre-Training with Efficient Audio Transformer. arXiv preprint. 2024. URL: https://arxiv.org/abs/2401.03497

10. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. arXiv preprint. 2019. URL: https://arxiv.org/abs/1912.10211

Abstract. *The paper examines the current state and possibilities of processing information flows in modern robotic systems with artificial intelligence tools. Comparative characteristics of components of various systems and technologies are presented to assess the possibilities of their use in different conditions.*

Key words: *artificial intelligence, robotic systems, sound information, information processing, cloud service, mathematical models.*

Статтю надіслано: 26.01.2026