АКТУАЛЬНІ ПИТАННЯ ТЕХНІЧНИХ НАУК

УДК 004.93:681.586
DOI https://doi.org/10.32782/2663-5941/2025.4.2/44

Вергелес А.В.
Національний університет «Львівська політехніка»

Юрчак І.Ю.
Національний університет «Львівська політехніка».
ORCID ID: 0009-0005-9100-8511

ЛОКАЛЬНА СИСТЕМА ВИЯВЛЕННЯ СЛОВА-ТРИГЕРА НА БАЗІ ESP32 ДЛЯ ГОЛОСОВОГО КЕРУВАННЯ БЕЗ ВИКОРИСТАННЯ ХМАРНИХ СЕРВІСІВ

Дослідження спрямоване на розроблення та експериментальне обґрунтування локальної системи виявлення слова-тригера на основі мікроконтролера ESP32, орієнтованої на автономне голосове керування вбудованими пристроями без залучення хмарних сервісів. Проблематика обумовлена зростанням поширення голосових інтерфейсів у побутовій електроніці, робототехніці та мобільних сенсорних системах, де стабільність роботи та конфіденційність голосових даних мають вирішальне значення. Використання хмарних підсистем розпізнавання мовлення супроводжується залежністю від мережевого каналу, збільшенням латентності та ризиком передачі приватних акустичних даних третім сторонам. У зв'язку з цим актуальним є створення методів виявлення мовного тригера без винесення обчислень за межі пристрою.

У роботі запропоновано модель, побудовану на компактній згортковій нейронній мережі, що оперує мел-спектрограмним представленням аудіосигналу та оптимізована для виконання у середовищі з обмеженими обчислювальними ресурсами. Застосовано повну квантизацію параметрів моделі до формату int8 та статичне планування пам'яті відповідно до парадигми TensorFlow Lite Micro, що дозволило досягти передбачуваного часу обчислення інференсу у режимі безперервного фонового прослуховування. Навчання здійснено на підмножині корпусу Google Speech Commands із формуванням збалансованих класів тригерного та нетригерного мовлення, включно з побутовими шумами для зменшення ймовірності хибних активацій.

Експериментальна оцінка засвідчила стабільність розпізнавання слова-тригера у діапазоні акустичних умов, характерних для внутрішніх приміщень, а час реакції системи перебував у межах 120–190 мс, що відповідає вимогам режиму реального часу. Отримані результати підтверджують можливість використання ESP32 як апаратної платформи для створення конфіденційних, енергоефективних та автономних голосових інтерфейсів без залучення зовнішніх серверних обчислювальних ресурсів. Подальші дослідження можуть бути спрямовані на розширення словника команд та впровадження адаптивних механізмів шумостійкості.

Ключові слова: *голосове керування, виявлення слова-тригера, ESP32, локальна обробка мовлення, згорткова нейронна мережа, квантизація моделі, TensorFlow Lite Micro.*

Актуальність проблеми. Застосування голосового керування у побутових та вбудованих системах зростає, однак більшість рішень для розпізнавання мовних команд покладаються на хмарні сервіси. Це призводить до залежності від стабільності мережевого з'єднання, збільшення затримок реакції та виникнення ризиків розкриття приватних голосових даних. Для пристроїв із обмеженими обчислювальними ресурсами характерною є потреба у безперервному фоновому прослуховуванні при мінімальному енергоспоживанні та гарантованому часі обробки сигналу.

Тому актуальним є розроблення локальної системи виявлення слова-тригера, здатної працювати автономно без передачі аудіосигналу на зовнішні сервери.

Додатковим чинником актуальності є розширення сфери застосування мікроконтролерних платформ у «розумних» пристроях, побутових сенсорних вузлах, робототехніці, системах безпеки та медичних асистивних засобах. У таких сценаріях необхідні низька латентність, передбачуваний час відгуку та стабільність роботи у змінних акустичних умовах. Використання локальних методів виявлення тригерного слова дозволяє забезпечити керування в реальному часі без підключення до мережі та зберегти конфіденційність голосового каналу, що визначає практичну значущість дослідження.

Аналіз останніх досліджень та публікацій. Актуальність задачі виявлення слова-тригера зумовлена потребою забезпечення локального голосового керування на пристроях із обмеженими апаратними ресурсами та вимогами гарантованого часу відгуку. Дослідження в цій сфері зосереджені на пошуку компактних та енергоефективних моделей, здатних працювати в автономному режимі без передачі мовного сигналу до хмарної інфраструктури. Особлива увага приділяється забезпеченню стійкості до фонового шуму та зменшенню ймовірності хибних активацій.

У роботі Дж. Бушур та Ч. Чен «Дослідження нейронних мереж для виявлення ключових слів на периферійних пристроях» (2023) [1] розглядаються особливості вибору архітектур нейронних мереж для режиму постійного фонового прослуховування на крайових пристроях. Автори досліджують компроміси між точністю класифікації та обчислювальною вартістю, показуючи переваги компактних згорткових моделей при роботі з мелспектрограмами. Наголошується, що ключовим фактором є зменшення кількості параметрів при збереженні стабільності розпізнавання.

У публікації Р. Девіда та співавт. «TensorFlow Lite Micro: вбудоване машинне навчання для систем TinyML» (2021) [2] описано підхід до виконання інференсу на мікроконтролерах без використання динамічного керування пам'яттю. Запропонована модель виконання базується на статичній виділеній обчислювальній арені, що унеможливлює фрагментацію пам'яті та забезпечує передбачуваний час відгуку критичну характеристику для систем із безперервним акустичним моніторингом.

Стаття Й. Свірського, У. Шахама та О. Лінденбаума «Розріджена бінаризація для швидкого пошуку ключових слів» (2024) [3] пропонує використання розріджених бінаризованих спектральних представлень для підвищення швидкодії та шумостійкості KWS-моделей. Використання бінарних масок дозволяє зменшити обчислювальну складність без суттєвого падіння точності, що робить підхід перспективним для реалізації на пристроях з обмеженими обчислювальними можливостями.

Дослідження підтверджують ефективність компактних згорткових моделей, квантизації параметрів та статичного планування пам'яті для локального виявлення слова-тригера на мікроконтролерах. Представлені підходи закладають методологічну основу для розробки автономних голосових інтерфейсів, орієнтованих на низьке споживання ресурсів і стабільний час реакції системи.

Метою дослідження є розроблення та експериментальна перевірка локальної системи виявлення слова-тригера, здатної працювати в реальному часі на мікроконтролері ESP32 без залучення хмарних сервісів. Завданням є побудова компактної моделі розпізнавання на основі CNN (Згорткова нейронна мережа, Convolutional neural network) з подальшою квантизацією та інтеграцією у прошивку, забезпечення стабільного часу інференсу та низького рівня хибних спрацювань при обмеженому обсязі пам'яті та обчислювальних ресурсів. Особлива увага приділяється оцінюванню стійкості системи до фонових акустичних завад та її працездатності в автономному режимі постійного прослуховування.

Виклад основного матеріалу. Голосове керування дедалі частіше застосовується в побутових і вбудованих пристроях, однак значна частина таких систем базується на використанні хмарних сервісів для розпізнавання аудіосигналів. Це створює залежність від якості мережевого з'єднання, збільшує затримку відповіді та ставить під питання конфіденційність переданих даних. Одним із підходів до усунення цих обмежень є локальне визначення слова-тригера без звернення до зовнішніх обчислювальних ресурсів. У цьому дослідженні розглядається побудова такої системи на мікроконтролері з обмеженими апаратними можливостями, що дозволяє забезпечити автономну роботу та відгук у реальному часі.

У запропонованій системі не виконується повне розпізнавання мовного сигналу, а визначається лише момент вимови слова-тригера. Такий

підхід суттєво зменшує обчислювальні витрати та дозволяє реалізувати алгоритм на мікроконтролері з обмеженими ресурсами. Обробка сигналу здійснюється послідовно, починаючи із захоплення аудіофрагмента з мікрофона та його нормалізації, після чого формується мел-спектрограма як компактне спектральне подання мовного сигналу, узагальнена схема наведена на Рис. 1. Цей спектральний образ подається на CNN, яка виконує класифікацію та визначає, чи присутнє у фрагменті слово-тригер. Результат класифікації використовується як сигнал активації подальшої логіки керування.

Для навчання моделі використано підмножину датасету Google Speech Commands [4], що містить одно секундні фрагменти мовлення, записані при частоті дискретизації 16 кГц. Цей набір даних є репрезентативним для завдання локального визначення слова-тригера на малопотужних пристроях, оскільки охоплює варіативність вимови, тембру та рівня шуму середовища. У дослідженні увага була спрямована на два класи: цільове слово-тригер та узагальнений клас «не тригер», що включав довільні мовні та немовні сигнали. Така двокласова схема дозволяє суттєво зменшити складність моделі та стабілізувати час інференсу на мікроконтролері, що є критичним для систем реального часу.

Для підвищення стійкості було сформовано розширений фоновий підклас, у який увійшли побутові шуми, короткі уривки мови інших осіб, паузи та шурхоти. Така стратегія знижує частоту хибних активацій у ситуаціях із підвищеним рівнем акустичних завад, що підтверджується попередніми дослідженнями з ключового розпізнавання на edge-пристроях. Вибірка була поділена на тренувальну (80%), валідаційну (10%) та тестову (10%) підмножини (Рис. 2) із забезпеченням рівномірного представлення голосів для запобігання переадаптації під конкретні акустичні профілі.

Аудіосигнал надходить із мікрофона I²S, оцифрованого у форматі 16-bit PCM при частоті 16 кГц. Первинний сигнал нормалізується для зменшення залежності від гучності мовця. Далі застосовується віконне перетворення із розміром вікна 25 мс та кроком 10 мс, що забезпечує достатню часову роздільність при збереженні помірної обчислювальної складності. На спектральному рівні використано мел-спектрограму з 40 фільтрами, що дозволяє інтерпретувати акустичні дані як двовимірну матрицю інтенсивностей (аналог зображення), оптимальну для CNN.

Цей підхід відповідає сучасним практикам застосування CNN для виявлення ключового слова у потоці мовлення, оскільки мел-спектрограма усуває зайву часову нестаціонарність та підкреслює діапазони частот, релевантні для артикуляційних характеристик мовлення. Додатково було перевірено можливість використання MFCC, однак обчислення коефіцієнтів у реальному часі вимагало б більших ресурсів DSP-підсистеми ESP32, що суперечить принципам енергетичної ефективності.

Попередня обробка та інференс виконуються без використання динамічної пам'яті, відповідно до архітектури TensorFlow Lite Micro [5], що застосовує статичне виділення обчислювальної арени та не покладається на операційну систему або апаратну підтримку плаваючої коми. Така організація забезпечує відсутність коливань часу відповіді та робить систему передбачуваною для сценаріїв завжди-увімкнутого акустичного моніторингу.

Вибір моделі зумовлений необхідністю забезпечити стабільний інференс на обмежених обчислювальних ресурсах. CNN малої глибини є оптимальною для задачі виявлення слова-тригера, оскільки працює без накопичення часової пам'яті та ефективно обробляє мел-спектрограму як двовимірний сигнал. На відміну від рекурентних та трансформерних архітектур, CNN забезпечує фіксований час обчислення та не створює додаткових затримок, що дозволяє підтримувати безперервний режим прослуховування в реальному часі.

Модель включає два згорткові шари з активацією ReLU та операціями max-pooling, після яких використовується один щільний вихідний шар.

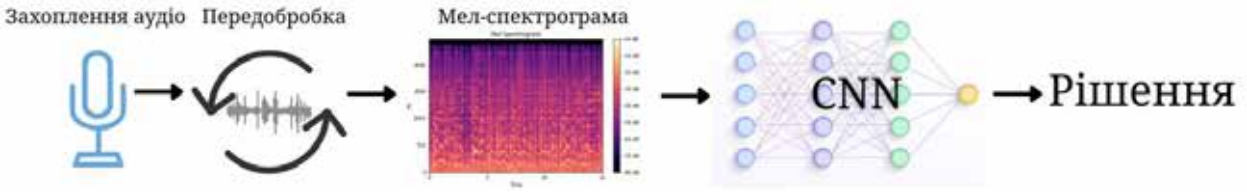

Рис. 1. Узагальнена схема обробки сигналу під час виявлення слова-тригера

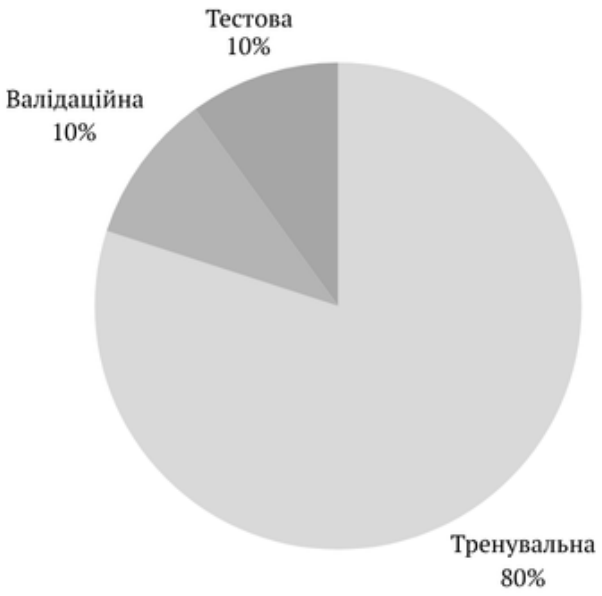

Рис. 2. Кругова діаграма розподілу вибірки

Така конфігурація забезпечує достатню виразність ознак при мінімальній кількості параметрів. Розмір моделі підібрано так, щоб вона повністю розміщувалась у flash-пам'яті ESP32 та виконувала інференс у реальному часі без перевищення доступного обсягу оперативної пам'яті.

Для зменшення ресурсоємності модель була повністю квантизована з float32 до int8. Це зменшило розмір ваг та активацій у пам'яті та дозволило виконувати інференс на цілочисельних операціях без використання плаваючої коми. Обчислювальна арена виділяється статично, що гарантує фіксований час виконання та стабільну затримку під час безперервного прослуховування.

Сконвертована модель у форматі TensorFlow Lite інтегрується у прошивку ESP32 у вигляді масиву байтів, що розміщується у flash-пам'яті. Під час роботи пристрою аудіосигнал буферизується з фіксованим ковзним вікном, формується спектрограма, після чого виконуються послідовні виклики інференсу. Організація циклу обробки забезпечує синхронізацію між надходженням аудіо та обчислювальними блоками без накопичення небажаної затримки.

Оцінювання моделі включає точність класифікації, показники precision і recall для події активації, а також частоту хибних спрацювань у різних акустичних умовах. Додатково вимірюються затримка інференсу та витрати ресурсів: обсяг flash-пам'яті під модель, використання оперативної пам'яті буфером та навантаження ядра під час безперервної обробки. Ці метрики визначають придатність системи до автономної роботи в режимі постійного прослуховування.

Для експериментальної перевірки система тестувалася в різних акустичних умовах та при зміні фонового навантаження сигналом. Оцінювалися стійкість до шумів, чутливість до вимови тригерного слова та час реакції на подію активації. Умови тестування відповідали режиму нормальної експлуатації: приміщення середнього об'єму, відстань від мовця до мікрофона 40–60 см, робота в автономному режимі без зовнішнього серверного процесингу. Результати наведено у таблицях 1-2.

Таблиця 1

Стійкість до акустичних завад

Акустичні умови (рівень звукового тиску, дБ)	Характер роботи детектора слова-тригера
30–40	стабільне виявлення цільового сигналу
45–55	періодичне повторне подання тригера для підтвердження
60+	виявлення можливе за умови чіткої артикуляції

Таблиця 2

Темпоральні характеристики відгуку

Умови оточення	Середній час від вимовлення слова-тригера до реакції системи
низький рівень акустичного фону	120–150 мс
присутність фонової мовної активності	150–190 мс

Отримані дані показують, що система зберігає стабільне виявлення тригерного слова за умов низького та помірного фонового шуму, а збільшення рівня акустичних завад переважно позначається на потребі чіткішої артикуляції. Час реакції залишається у межах 120–190 мс залежно від фонової мовної активності, що відповідає вимогам режиму реального часу. Сукупно це свідчить про можливість тривалої автономної роботи детектора без залучення зовнішніх обчислювальних ресурсів.

Висновки. Дослідження показало можливість реалізації локального виявлення слова-тригера на мікроконтролері ESP32 без використання зовнішніх серверних ресурсів. Застосування компактної CNN з подальшою повною квантизацією дозволило розмістити модель у flash-пам'яті обсягом менше 250 кБ та забезпечити стабільний час інференсу в межах 100 мс. Система зберігає працез-

датність у режимі постійного прослуховування аудіосигналу та демонструє середню точність виявлення тригерного слова на рівні близько 95 % за умов побутового фонового шуму. Експериментальні вимірювання підтвердили прийнятну стійкість до акустичних завад і відсутність суттєвих флуктуацій затримки відгуку.

Отримані результати свідчать про можливість використання мікроконтролера ESP32 як апаратної основи для автономних голосових інтерфейсів у вбудованих системах і пристроях з обмеженими ресурсами. Подальший розвиток роботи може бути спрямований на удосконалення алгоритмів приглушення шумів, адаптивне налаштування порогів активації та розширення набору розпізнаваних голосових команд без суттєвого збільшення обчислювального навантаження.

Список літератури:

1. Bushur J., Chen C. Neural network exploration for keyword spotting on edge devices. *Future internet*. 2023. Vol. 15, no. 6. P. 219. DOI: https://doi.org/10.3390/fi15060219.

2. David R., Duke J., Jain A., Reddi V. J., Jeffries N., Li J., Kreeger N., Nappier I., Natraj M., Regev S., Rhodes G., Warden P. TensorFlow Lite Micro: Embedded Machine Learning for TinyML Systems. *Proceedings of Machine Learning and Systems.* 2021. Vol. 3. P. 800–811.

3. Svirsky J., Shaham U., Lindenbaum O. Sparse Binarization for Fast Keyword Spotting. *Proceedings of Interspeech.* 2024. P. 3010–3014. DOI: https://doi.org/10.21437/Interspeech.2024-389.

4. Google/speech_commands · datasets at hugging face. *Hugging Face – The AI community building the future.* URL: https://huggingface.co/datasets/google/speech_commands (date of access: 18.07.2025).

5. GitHub – tensorflow/tflite-micro: infrastructure to enable deployment of ML models to low-power resource-constrained embedded targets (including microcontrollers and digital signal processors). *GitHub*. URL: https://github.com/tensorflow/tflite-micro (date of access: 17.07.2025).

Verheles A.V., Yurchak I.Yu. LOCAL TRIGGER WORD DETECTION SYSTEM BASED ON ESP32 FOR VOICE CONTROL WITHOUT CLOUD SERVICES

The research is focused on the development and experimental validation of a local trigger word detection system based on the ESP32 microcontroller, intended for autonomous voice control of embedded devices without the use of cloud services. The relevance of this work is determined by the growing dissemination of voice interfaces in consumer electronics, robotics, and mobile sensor systems, where operational stability and the confidentiality of voice data are critical. The use of cloud-based speech recognition introduces dependence on network connectivity, increases response latency, and poses risks associated with transferring private acoustic data to external servers. Therefore, creating trigger word detection methods that operate entirely on-device is an urgent task.

The proposed approach utilizes a compact convolutional neural network operating on mel-spectrogram representations of the speech signal and optimized for execution under limited computational resources. Full quantization of model parameters to the int8 format and static memory allocation according to the TensorFlow Lite Micro paradigm were applied, ensuring predictable inference time in continuous background listening mode. The model was trained on a selected subset of the Google Speech Commands dataset with balanced trigger and non-trigger classes, including environmental noise samples to reduce false activation rates.

Experimental evaluation confirmed stable trigger word recognition performance under typical indoor acoustic conditions, with system reaction times ranging from 120 to 190 ms, which meets real-time requirements. The obtained results demonstrate the feasibility of using ESP32 as a hardware platform for confidential, energy-efficient, and autonomous voice interfaces without reliance on external computational infrastructure. Future work may focus on extending the command vocabulary and integrating adaptive noise resilience mechanisms.

Key words: *voice control, trigger word detection, ESP32, on-device speech processing, convolutional neural network, model quantization, TensorFlow Lite Micro.*

Дата надходження статті: 14.07.2025
Дата прийняття статті: 21.07.2025
Опубліковано: 27.10.2025