УДК 004.8:004.934

ГІБРИДНА СИСТЕМА ДЕТЕКЦІЇ СИНТЕЗОВАНОГО МОВЛЕННЯ НА ОСНОВІ SSL-ЕМБЕДДІНГІВ

Ростовцев В.В., Гриньов С.А.
e-mail: volodymyr.rostovtsev@nure.ua, serhii.hrynov@nure.ua
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This paper presents a hybrid system for synthetic speech detection that integrates spectral, phase-based, and prosodic acoustic features with self-supervised learning (SSL) embeddings extracted from the WavLM model. The proposed multimodal approach combines low-level handcrafted features (LFCC, jitter, shimmer, phase stability) with high-level context-aware representations derived from intermediate transformer layers. Such integration enables the model to capture both fine-grained acoustic artifacts and global structural inconsistencies in synthetic speech. The proposed architecture shows strong potential for applications in biometric security and digital media integrity.

Швидкий прогрес у галузі глибокого навчання призвів до появи генеративних моделей, здатних створювати аудіоконтент, який майже неможливо відрізнити від справжнього людського голосу. Технології нейронного синтезу мовлення та клонування голосу створюють нові загрози для систем біометричної автентифікації та інформаційної безпеки. Проблема автоматичного виявлення штучно згенерованого голосу стає критичною у контексті протидії дезінформації та шахрайству. Метою даної роботи є розробка та дослідження гібридної системи детекції, що базується на аналізі низькорівневих акустичних артефактів та високорівневих представлень, отриманих за допомогою самокерованого навчання.

Сучасні TTS-архітектури зазвичай використовують двоетапну схему генерації проміжного представлення (мел-спектрограми) та подальшого синтезу хвильової форми за допомогою вокодера. Найбільш розповсюджені вокодери (HiFi-GAN та MelGAN), які побудовані на основі генеративно-змагальних мереж (GAN). Вони здатні створювати високоякісний сигнал, проте часто припускаються помилок у відтворенні фазової структури та високочастотних гармонік [1].

Більш складними для детекції є end-to-end моделі типу VITS, у яких відсутній явний етап генерації спектрограми, що зменшує кількість класичних артефактів. Окрему категорію становлять комерційні системи (ElevenLabs), які використовують масштабні трансформерні архітектури. Це забезпечує високу природність мовлення, зокрема у плані просодики та емоційного забарвлення.

Для підвищення ефективності детекції запропоновано використовувати комбінацію ознак на різних рівнях абстракції.

Рівень 1. Спектральні ознаки. Артефакти нейронних вокодерів часто проявляються у високочастотній області спектра. Використання лінійних частотних кепстральних коефіцієнтів (LFCC) замість традиційних MFCC дозволяє уникнути стискання високочастотної зони, характерного для мел-шкали, та ефективніше фіксувати цифрові аномалії.

Рівень 2. Просодичні характеристики. Аналіз мікроколивань основної частоти (F0), а також показників Jitter (варіативність періоду) та Shimmer (варіативність амплітуди) дозволяє оцінити природність мікроваріацій мовлення. У природному голосі ці показники мають хаотичну природу, зумовлену фізіологією гортані, тоді як синтетичні моделі часто генерують занадто «гладкі» або циклічні переходи.

Рівень 3. Фазові характеристики. Зазвичай, сучасні вокодери фокусуються на відтворенні амплітудного спектра, приділяючи менше уваги фазової відповідності. Аналіз миттєвої фази сигналу (фазова стабільність) дозволяє виявити розриви та нелінійності, характерні для штучно згенерованого звуку.

Візуалізація сигналів у формі спектрограм (рис. 1) підтверджує наявність артефактів, що лишаються поза увагою нейронних вокодерів.

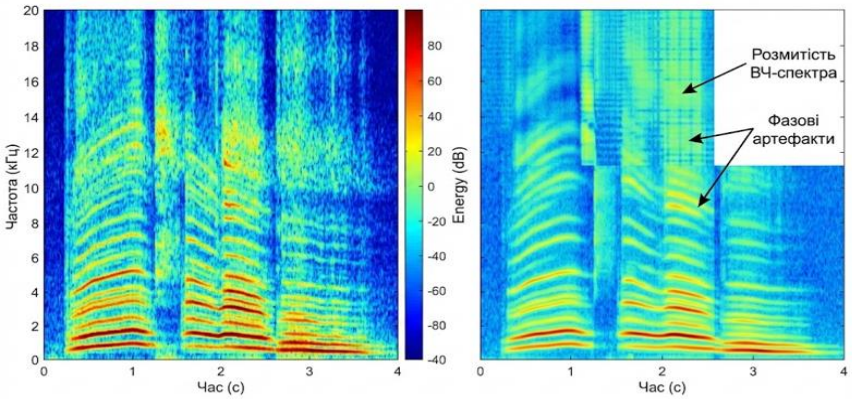

Рисунок 1 – Порівняльна діаграма спектрограм

На спектрограмі природного мовлення (рис. 1а) спостерігається чітка структура нижніх гармонік та природний спектральний нахил з наявністю стохастичного шуму на високих частотах. Синтезований голос (рис. 1б) містить специфічні «цифрові сліди»: «шахові» артефакти (Checkerboard artifacts) – періодичні сплески енергії, спричинені операціями деконволюції; надмірне згладжування (Over-smoothing) – занадто «чистий» спектр у верхніх діапазонах; фазові розриви – різкі вертикальні лінії на межах синтезованих кадрів. Математично ці аномалії фіксуються через LFCC, дозволяючи класифікатору виявляти специфічні «цифрові відбитки». Таким чином, візуальний аналіз підтверджує, що сучасні системи клонування голосу не досягають повної ідентичності з біологічним сигналом.

Окрім низькорівневих характеристик, у роботі застосовується інтеграція моделей самокерованого навчання (SSL), які формують узагальнені представлення аудіосигналу. Використовуються моделі wav2vec 2.0 та WavLM, навчені на великих обсягах немаркованого мовлення [2]. Для задачі детекції застосовуються представлення середніх шарів трансформера, що поєднують акустичні та контекстні характеристики сигналу. Модель WavLM також враховує зашумлені й перекриті сигнали, що підвищує стійкість системи до реальних умов запису. Це покращує здатність системи до узагальнення та виявлення синтетичного мовлення, створеного невідомими моделями.

Гібридна інтеграція ознак здійснюється у спільному векторному просторі. Низькорівневі спектральні, фазові та просодичні параметри після нормалізації проходять через блок проєкції для узгодження розмірності. SSL-ембеддинги обробляються окремим шаром проєкції. Після вирівнювання розмірності ознаки об'єднуються шляхом конкатенації у спільний вектор представлення. Далі застосовується механізм уваги (attention-based fusion), який адаптивно зважує внесок окремих груп ознак. Отримане представлення подається до багатошарового перцептрона (MLP), який виконує бінарну класифікацію [3].

Для оцінювання системи використовується набір даних ASVspoof 2019 (LA subset). Основні метрики – EER (Equal Error Rate, $P_{\{fa\}(\theta)} = P_{\{\text{miss}\}(\theta)}$, де $P_{\{fa\}}$ – ймовірність хибної тривоги (False Acceptance), а $P_{\{\text{miss}\}}$ – ймовірність пропуску цілі (False Rejection) при певному порозі θ; AUC ROC – площа під кривою помилок, що демонструє загальну роздільну здатність моделі.

У роботі обґрунтовано мультимодальний підхід до детекції ШІ-мовлення. Поєднання класичної цифрової обробки сигналів (аналіз фази та спектра) із сучасними нейромережевими архітектурами (WavLM) забезпечує високу точність верифікації контенту.

Список використаних джерел:

1. Kong J., Kim J., Bae J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis.URL:https://arxiv.org/abs/2010.05646.

2. Chen S., et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. IEEE Journal of Selected Topics in Signal Processing. 2022. Vol. 16. P. 1505–1518. URL: https://arxiv.org/abs/2110.13900.

3. Fast neural network and its adaptive learning in classification problems / Y. V. Bodyanskiy et al. Radio electronics, computer science, control. 2025. No. 3. P. 45–51. DOI: https://doi.org/10.15588/1607-3274-2025-3-5.