Сєрих С.О.,
кандидат технічних наук, доцент кафедри Комп'ютерних наук, Державний університет інформаційно-комунікаційних технологій
Попов А.О.,
аспірант кафедри Комп'ютерних наук, Державний університет інформаційно-комунікаційних технологій
(*м.Київ, Україна*)

МЕТОДИ ПРОТИДІЇ ПРОВЕДЕННЮ ІНФОРМАЦІЙНИХ ОПЕРАЦІЙ З ВИКОРИСТАННЯМ НЕЙРОННИХ МЕРЕЖ СИНТЕЗУ МОВЛЕННЯ

У статті розглянуто можливості використання нейронних мереж генерації мовлення для проведення інформаційних операцій. Проаналізовано методи протидії таким системам.

Постановка задачі. Необхідно провести аналіз наявних систем виявлення синтезованого мовлення, проаналізувати існуючі методи протидії. Визначити ефективність існуючих інструментів та запропонувати можливі шляхи розвитку.

Мета дослідження. Визначити інструменти та підходи які можна використати для протидії використання нейронних мереж синтезу мовлення для проведенні інформаційних операцій.

Результати дослідження. Зараз людство знаходиться на найвищому за всю історії рівні свободи доступу до інформації, завдяки інтернету будь хто може спілкуватися і отримувати інформацію з будь якого куточку планети, що є величезним благом. Однак такий рівень розвитку людства несе і нові виклики, оскільки люди не перестали вести війни, вони як і раніше періодично трапляються зараз, невідмінно від війн двадцятого сторіччя де доступ до інформації був сильно обмежений технологічним розвитком людства, зараз надзвичайно просто провести інформаційну по дезінформації населення ворожої країни. Нерідко такі операції проводяться з використанням генерованого мовлення яке копіює голос відомих спікерів країни на які направленні ці операції по дезінформації, оскільки часто знайомі голоси високо посадовців сприймаються як перевірені джерела інформації.

Сучасні методи визначення поєднують класичні спектральні ознаки (LFCC, MFCC, Mel-спектр) з SSL-фронтендами та трансформерними архітектурами над спектрограмою, важливу роль відіграє також агресивна аугментація даних і ансамблі моделей для підвищення стійкості до нових

генераторів. Для перевірки та порівняння рішень дослідники радять використовувати стандартизовані бенчмарки (ASVspoof, новіші датасети) і фреймворки оцінки, бо без узгодженого тестування, модель, що добре працює на одних генераторах, часто провалюється на інших. Цей підхід допомагає будувати автоматичні фільтри, які попередньо маркують підозрілі аудіо файли.

Можливе використання стандартів для документування походження контенту (content credentials, C2PA) і систем «контент-підпису» дозволяє офіційним джерелам підписувати аудіо файли (це може бути криптографічний підпис або зашифровані метадані), що дає змогу автоматично перевірити, чи був файл створений і/або відредагований авторизованим каналом. Водночас варто поєднувати зовнішній підпис з вбудованими водяними знаками, які витримують конвертацію форматів і відтворення через різні платформи — це ускладнить підробку офіційних записів. Для екосистеми медіа необхідно впровадити процес підпису й перевірки контенту в процеси створення та публікації інформації.

Технічно можливо використовувати нейронні водяні знаки, які прямо вбудовують «сигнатуру» оригінального мовлення у хвилю — при атаці або зміні голосу ця сигнатура дозволяє відновити ідентичність джерела або виявити маніпуляцію.

Висновки та перспективи. Розвиток технологій синтезу мовлення, що здатні відтворювати голоси із високим ступенем правдоподібності, створює серйозні виклики для інформаційної безпеки. У сучасних умовах війн і гібридних конфліктів зловмисники можуть використовувати такі інструменти для дезінформації, маніпуляції громадською думкою та підриву довіри до офіційних джерел інформації та влади, спрямованих на дестабілізацію держави. Традиційні методи виявлення виявляються недостатніми, тому наукова спільнота розробляє нові підходи, що поєднують класичні спектральні ознаки з сучасними нейромережевими архітектурами. Важливим напрямом стає створення систем цифрової аутентифікації та підпису контенту, які дозволяють відрізнити справжні повідомлення від штучно згенерованих.

Список використаних джерел:

1. Warning: Humans Cannot Reliably Detect Speech Deepfakes URL: https://arxiv.org/abs/2301.07829
2. Deepfake audio detection with spectral features and ResNeXt-based architecture. URL: https://surl.lt/sptaui
3. Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead.URL: https://surl.li/dntabb