КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (26), 2024
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2024.26.670
УДК 004.62

Савкова Тетяна Юріївна
студентка кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID ID: 0009-0009-1112-9508
tetiana.savkova.kb.2021@lpnu.ua

Опірський Іван Романович
д.т.н., професор, завідувач кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID ID: 0000-0002-8461-8996
ivan.r.opirskyi@lpnu.ua

Сабодашко Дмитро Володимирович
PhD, старший викладач кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID ID: 0000-0003-1675-0976
dmytro.v.sabodashko@lpnu.ua

ДОСЛІДЖЕННЯ СТІЙКОСТІ БІОМЕТРИЧНИХ СИСТЕМ АВТЕНТИФІКАЦІЇ ДО АТАК ІЗ ЗАСТОСУВАННЯМ ТЕХНОЛОГІЇ КЛОНУВАННЯ ГОЛОСУ НА ОСНОВІ ГЛИБИННИХ НЕЙРОННИХ МЕРЕЖ

Анотація. З розвитком технологій синтезу голосу на основі глибинних нейронних мереж зросли загрози, пов'язані з безпекою біометричних систем автентифікації, які використовують розпізнавання голосу. Ці системи, які вважалися надійними, можуть бути легко скомпрометовані через підроблені голоси, що створюються за допомогою передових моделей, таких як WaveNet, Tacotron 2 та інші сучасні алгоритми. В умовах сучасної кібербезпеки такі атаки ставлять під загрозу конфіденційність персональних даних, що викликає потребу у вдосконаленні методів захисту. Метою цієї статті є дослідження стійкості біометричних систем автентифікації до атак із застосуванням технології клонування голосу, аналіз ефективності сучасних методів синтезу для обходу таких систем, а також порівняльний огляд різних підходів, що дозволяють захистити голосові біометричні дані. Розглядаються технології, що дозволяють створювати точні і реалістичні синтетичні голоси, а також методи виявлення та захисту від підроблених сигналів. Стаття також аналізує поточні вразливості голосових систем і пропонує стратегії для підвищення стійкості до подібних атак, забезпечуючи користувачам більшу безпеку та конфіденційність.

Ключові слова: клонування голосу; біометричні системи автентифікації; глибинні нейронні мережі; WaveNet; Tacotron 2; безпека; синтез голосу.

ВСТУП

Ідентифікація на основі біометричних даних — це засіб автоматичного розпізнавання особистості на базі унікальних фізичних або поведінкових параметрів [1]. Це один із найперспективніших напрямів захисту систем від несанкціонованих впливів. Однак, незважаючи на всю привабливість, даний підхід пов'язаний з низкою серйозних проблем. Спочатку розвиток та впровадження біометричних систем пов'язували зі статичними біометричними ознаками користувача (зображення обличчя, папілярний візерунок пальця та райдужна оболонка ока), які добре зарекомендували себе у криміналістиці. Однак, на цей час ці надії зруйновані, в першу чергу, через простоту

підробки [3]. Тому акцент робиться на сучасні біометричні системи автентифікації, зокрема ті, що базуються на розпізнаванні голосу, дедалі частіше застосовуються у різних сферах життя — від банківських установ до побутових пристроїв. Вони забезпечують зручність і безпеку, що дозволяє користувачам автентифікуватися без використання традиційних паролів. Проте, розвиток технологій глибинного навчання створив нові виклики для таких систем. Поява передових методів клонування голосу, що використовують глибинні нейронні мережі (далі ГНМ), таких як WaveNet і Tacotron 2, значно підвищила якість синтезованого мовлення, роблячи його майже невідрізним від оригінального. Це викликає серйозні побоювання щодо безпеки, оскільки зловмисники можуть використовувати підроблені голоси для обходу біометричних систем. Найбільшу загрозу для впровадження та використання біометричних систем створюють спуфінг-атаки — в контексті безпеки мережі, це випадок, коли особа або програма маскується під іншу за допомогою фальсифікації даних, і тим самим отримує незаконну перевагу [2].

ГНМ здатні моделювати найдрібніші деталі голосу, включаючи інтонації, тембр і ритм, що дозволяє створювати реалістичні аудіозаписи голосу цільової особи. Таким чином, ці системи стикаються з новими загрозами, що потребують розробки ефективних методів виявлення і протидії синтезованим голосам. Дослідження в цьому напрямку є критично важливим для забезпечення надійного захисту персональних даних та конфіденційності користувачів.

Постановка проблеми. Системи автентифікації на основі голосу вважалися ефективними та надійними для захисту інформації, проте з розвитком методів глибинного навчання з’явилися технології, які дозволяють точно імітувати голос користувача. Це створює значні ризики, оскільки синтезований голос можна використовувати для обману системи і отримання доступу до конфіденційних даних, фінансових ресурсів та інших критично важливих сервісів. Існуючі методи захисту часто не здатні відрізнити справжній голос від підробленого, що ставить під загрозу безпеку таких систем.

Отже, постає необхідність дослідження вразливостей голосових біометричних систем та розробки нових стратегій захисту.

Аналіз останніх досліджень та публікацій. Згідно з останніми дослідженнями, основною загрозою для біометричних систем автентифікації, що базуються на голосі, є використання сучасних моделей глибинного навчання для синтезу голосу [4], [5]. Розвиток таких моделей, як Tacotron 2, WaveNet, і VITS, зробив можливим створення реалістичних синтетичних голосів, які майже неможливо відрізнити від оригінальних. У дослідженні, представленому на конференції IberSPEECH 2022, дослідники підтвердили високу точність таких моделей та їх здатність до синтезу голосів з реалістичними тембрами та інтонаціями, що створює загрозу для систем автентифікації на основі голосу [7], [8].

Дослідження показало, що успішні атаки на системи автентифікації за допомогою синтетичного голосу можуть обійти існуючі механізми перевірки. Це підтверджує, що використання ГНМ для клонування голосу стає значною загрозою, оскільки навіть перевірені системи, як-от Deep Speaker, можуть бути скомпрометовані. Аналіз у роботах з Київського Політехнічного Інституту також підкреслює проблему безпеки таких систем, особливо якщо вони не мають додаткових методів перевірки для виявлення синтетичних голосів [9].

Статисчність синтезу голосу за останні кілька років значно покращилася. Наприклад, за даними дослідження з використанням моделей WaveNet, синтетичні голоси можуть успішно обманути системи автентифікації з ймовірністю до 85%. Подібні результати було продемонстровано в інших дослідженнях, що свідчить про значне

вдосконалення технологій клонування голосу. За результатами проведеного аналізу в роботі «Detection of Synthetic Speech Attacks», відзначено, що існує значна потреба у вдосконаленні методів ідентифікації підроблених голосів для посилення захисту голосових біометричних систем [7], [8].

Дослідження підкреслюють необхідність вдосконалення засобів безпеки, зокрема через впровадження нових методів, які зможуть розпізнавати синтетичні голоси та блокувати спроби обходу системи. Враховуючи швидкий розвиток технологій глибинного навчання, особливо моделей синтезу голосу, підвищення стійкості до таких загроз стає першочерговим завданням для розробників біометричних систем автентифікації.

Мета статті. Метою цієї статті є аналіз сучасних методів клонування голосу, що базуються на ГНМ, і оцінка їх ефективності для атак на біометричні системи автентифікації. На основі порівняння різних моделей пропонуються підходи до підвищення стійкості біометричних систем до атак такого типу.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Огляд технологій клонування голосу

Проблема клонування голосу стала вийшла на новий рівень завдяки зростанню популярності ГНМ. Синтез мовлення здебільшого базується на автоматичному перетворенні тексту в голос (Text-to-Speech, TTS). Справжній прорив у TTS стався у 2016 році, коли Google представила WaveNet, глибоку нейронну мережу, здатну генерувати більш природне та схоже на людину мовлення шляхом прямого моделювання необроблених звукових сигналів. Здатність WaveNet створювати мову, яка дуже нагадує якість людського голосу, встановила новий стандарт у цій галузі.

Після WaveNet, ще одним значним прогресом стало впровадження моделей трансформаторів у TTS. Трансформатори, спочатку розроблені для завдань обробки природної мови, спричинили революційні зміни в обробці довгострокових залежностей у мовленні, що призвело до покращення просодії (здатності системи синтезувати мову з природним ритмом і інтонацією, максимально наближеними до живої людської мови) та загальної якості мовлення. Це нововведення призвело до розробки таких моделей, як Tacotron 2 приблизно у 2020 році, які ще більше покращили природність і виразність синтетичного мовлення. Але про це згодом.

Такі системи як TTS використовують графеми — літери та їх комбінації, які транслітеруються у фонеми, найменші одиниці звуку. Основний ресурс цих систем — текст, а не сам звук. Синтез здійснюється шляхом перетворення тексту в еквівалентні слова, що включає нормалізацію тексту, фонетичну транскрипцію, групування в речення, а далі — конвертацію в звукові уявлення [6]. Якість оцінюється за схожістю з людським голосом і зрозумілістю мовлення [11] (рис. 1).

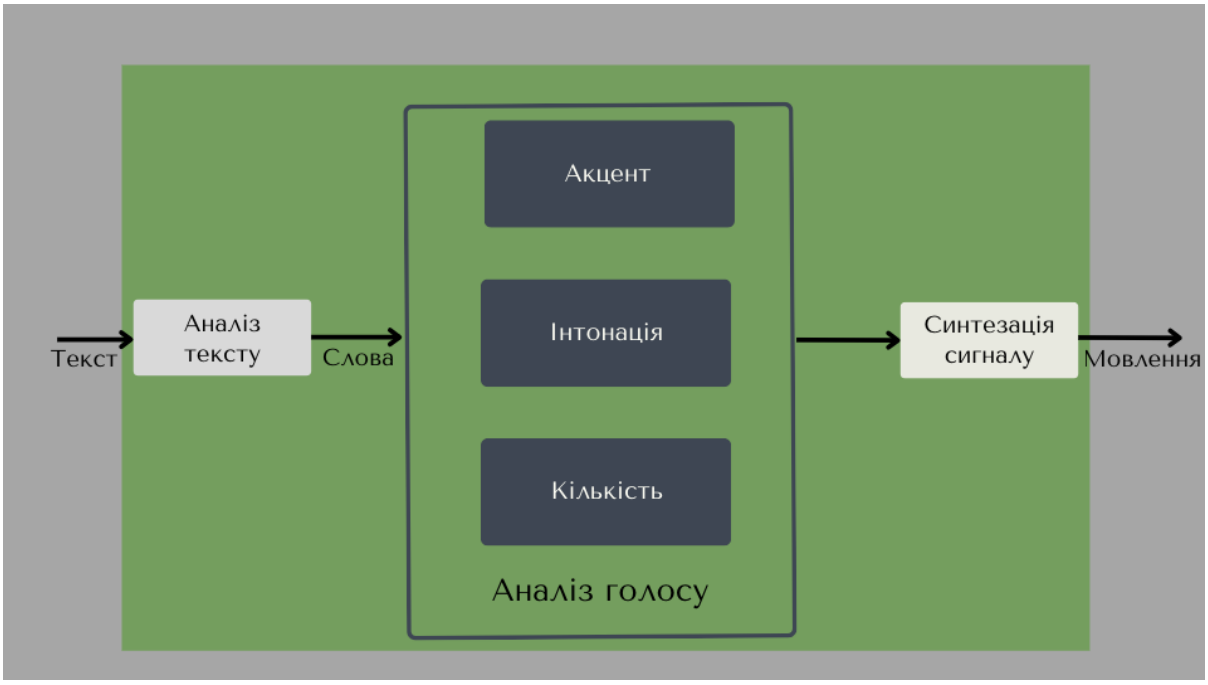

Рис. 1. Схема роботи типового синтезатора голосу TTS

Сучасні технології клонування голосу, зокрема WaveNet, Tacotron 2 та VITS, зробили суттєвий крок уперед у сфері синтезу мовлення, що має як позитивні, так і негативні наслідки для біометричних систем автентифікації. Ці моделі значно підвищили точність і реалістичність синтетичного мовлення, створюючи нові можливості для обману систем розпізнавання голосу.

WaveNet

WaveNet — це потужна нейронна мережа для генерації мовлення, розроблена компанією Google DeepMind. Вона використовує автогресивний підхід для створення високоякісного та натурального аудіосигналу, заснованого на аналізі попередніх семплів. WaveNet встановила нові стандарти у синтезі мовлення, однак має деякі обмеження, пов’язані з обчислювальною складністю та затримкою в реальному часі. Архітектура WaveNet наведена в табл. 1.

Таблиця 1

Архітектура WaveNet [8]

Компонент	**Опис**	**Функція**
Загальна структура	WaveNet складається з багатошарового підходу, де інформація передається через кілька рівнів обробки. Модель починається з початкових семплів, які підлягають обробці через серію згорткових шарів [7]	Обробка аудіосигналів та генерування звуку
Залишкові блоки	Основним елементом архітектури WaveNet є залишкові блоки, які дозволяють моделі вивчати більш складні представлення даних, не втрачаючи при цьому інформацію з попередніх шарів	Вивчення складних представлень даних
Вхідні дані	Аудіосигнал або спектрограма, що надходить на вхід	Введення інформації для обробки
Згорткові шари	Використовуються для витягування особливостей з вхідних даних. WaveNet використовує кілька згорткових шарів, які можуть мати різну ширину фільтра	Витягування ознак з аудіосигналу

Активаційні функції	Після кожного згорткового шару зазвичай застосовується функція активації ReLU (Rectified Linear Unit), яка допомагає моделі навчитися нелінійних відносин	Нелінійна обробка даних
Залишкове з'єднання	Залишкові з'єднання забезпечують короткі шляхи для передачі сигналу через мережу, зберігаючи важливу інформацію з попередніх шарів і запобігаючи затуханню градієнтів під час навчання	Підтримка інформації та уникнення затухання градієнтів
Об'єднання	Вихід залишкового блоку об'єднується з виходом наступного шару, що дозволяє моделі більш ефективно передавати інформацію і вивчати глибокі представлення аудіо	Покращення передачі інформації
Генерація аудіосигналу	WaveNet генерує аудіосигнал, прогнозуючи значення наступного семпла на основі вже згенерованих, створюючи контекстуальний зв'язок у часі	Створення високоякісного звуку
Модуляція часу	WaveNet включає механізми, що дозволяють модулювати час, забезпечуючи кращу синхронізацію генерації аудіосигналу з текстом або іншими сигналами	Синхронізація з текстом та іншими сигналами

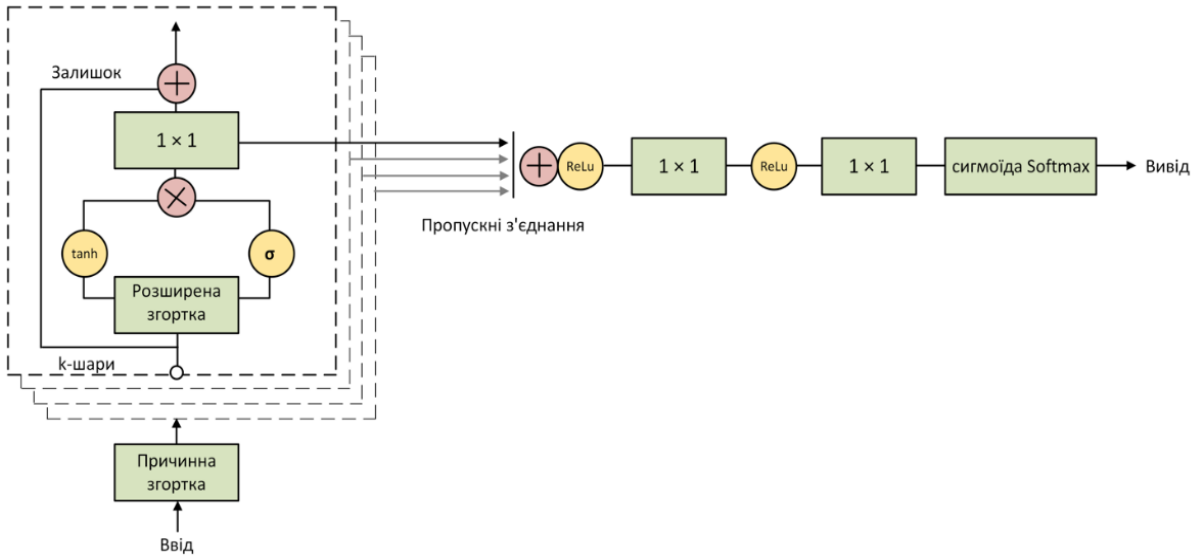

Рис. 2. Архітектурна модель WaveNet [9]

В даної моделі є дві основні переваги: висока реалістичність та гнучкість. **З**авдяки своїй архітектурі, WaveNet здатен генерувати звук, що максимально наближається до природного мовлення, з урахуванням усіх нюансів, таких як акценти та інтонації.

Також модель може бути налаштована для створення різних голосів та стилів мовлення, що робить її дуже універсальною.

Попри переваги також є і недоліки, а саме висока обчислювана вартість та затримка.

WaveNet є ресурсомісткою, оскільки вона повинна обробляти кожен семпл на основі всіх попередніх. Тобто для застосувань в реальному часі вона менш придатна. А процес генерації аудіосигналу може бути доволі затриманим, що ускладнює використання моделі в інтерактивних застосунках.

В цілому, WaveNet відкрила нові горизонти у технологіях синтезу мовлення, демонструючи потужність та можливості ГНМ. Її архітектура з використанням залишкових блоків забезпечує вивчення складних патернів у звукових даних, але потрібно враховувати також і труднощі зі обмеженнями швидкості та ресурсів, з якими доведеться стикнутись.

Tacotron 2

Tacotron 2, також розроблена Google — нейромережева архітектура для для синтезу мови безпосередньо з тексту. Tacotron 2 поєднує дві основні компоненти: перетворення тексту в спектрограму та перетворення спектрограми у звуковий сигнал через WaveNet. Завдяки такому підходу, Tacotron 2 може швидше генерувати голос, порівняно з автогресивним підходом WaveNet, при цьому зберігаючи високу якість звуку.

Використання цього компактного акустичного проміжного представлення дозволяє суттєво зменшити розмір архітектури WaveNet. Проте все ж можливі проблеми з інтонацією і артефакти в синтезованому голосі. Архітектура Tacotron 2 наведена в табл. 2 (також див. рис. 3).

Таблиця 2

Архітектура Tacotron 2

Компонент	**Короткий опис**	**Функція**
Загальна структура	Tacotron 2 поєднує два основні етапи: перетворення тексту в мел-спектрограму та перетворення спектрограми у звуковий сигнал через WaveNet-вокодер. Працює з нормалізованими текстовими даними та відповідними аудіосигналами, забезпечуючи природний звук	Швидка та ефективна генерація звуку з тексту з високою якістю
Блоки Tacotron 2	Основні компоненти, що виконують послідовні перетворення: з тексту в мел-спектрограму та зі спектрограми в звуковий сигнал. Використовує нейронні мережі та вдосконалені методи обробки для підвищення якості синтезу [7]	Обробка тексту та синтезування звуку з високою якістю та природністю
Текст у спектрограму	Застосовує рекурентні нейронні мережі (RNN) з механізмом уваги для перетворення тексту на мел-спектрограми. Забезпечує плавний перехід між текстом і аудіо	Витягування акустичних ознак з тексту, підготовка даних для WaveNet
Спектрограма в аудіо	Використовує модифікований WaveNet для перетворення мел-спектрограм на звуковий сигнал. Завдяки цьому підходу зберігається природна інтонація та висока якість аудіо	Синтезування звуку на основі спектрограм, що відповідає заданим текстовим даним
Використання уваги	Механізм уваги дозволяє моделі фокусуватися на різних частинах тексту під час генерації, забезпечуючи відповідність між текстом і аудіо, полегшує синтез довгих або складних речень	Поліпшення кореляції між текстом і згенерованим звуком, створення плавної та природної мови
Мел-спектрограми	Використовуються як акустичні проміжні ознаки, що дають можливість окремо навчати етапи обробки. Вони компактні та легко піддаються нейронній обробці, що зменшує складність процесу навчання	Підготовка акустичних ознак для покращення якості синтезованого звуку
Модифікований WaveNet	Генерує часові звукові хвилі на основі отриманих мел-спектрограм, використовуючи їх як умовний вхід. Це дозволяє створювати більш природний звук порівняно з традиційними підходами	Забезпечення реалістичного звуку з меншою кількістю артефактів, підвищення якості синтезу мовлення

ISSN 2663 - 4023

В даної моделі є три основні переваги. Модель проста, проте ефективна, адже завдяки компактній акустичній ознаці (мел-спектрограмі) можна окремо навчати два компоненти. А висока якість мовлення дозволяє за допомогою застосування WaveNet як вокодера створювати природні звуки без характерних артефактів, які були у попередніх методах. Також завдяки відносно спрощеній структурі модель може навчатися різним мовам і стилям мовлення, тобто є універсальною.

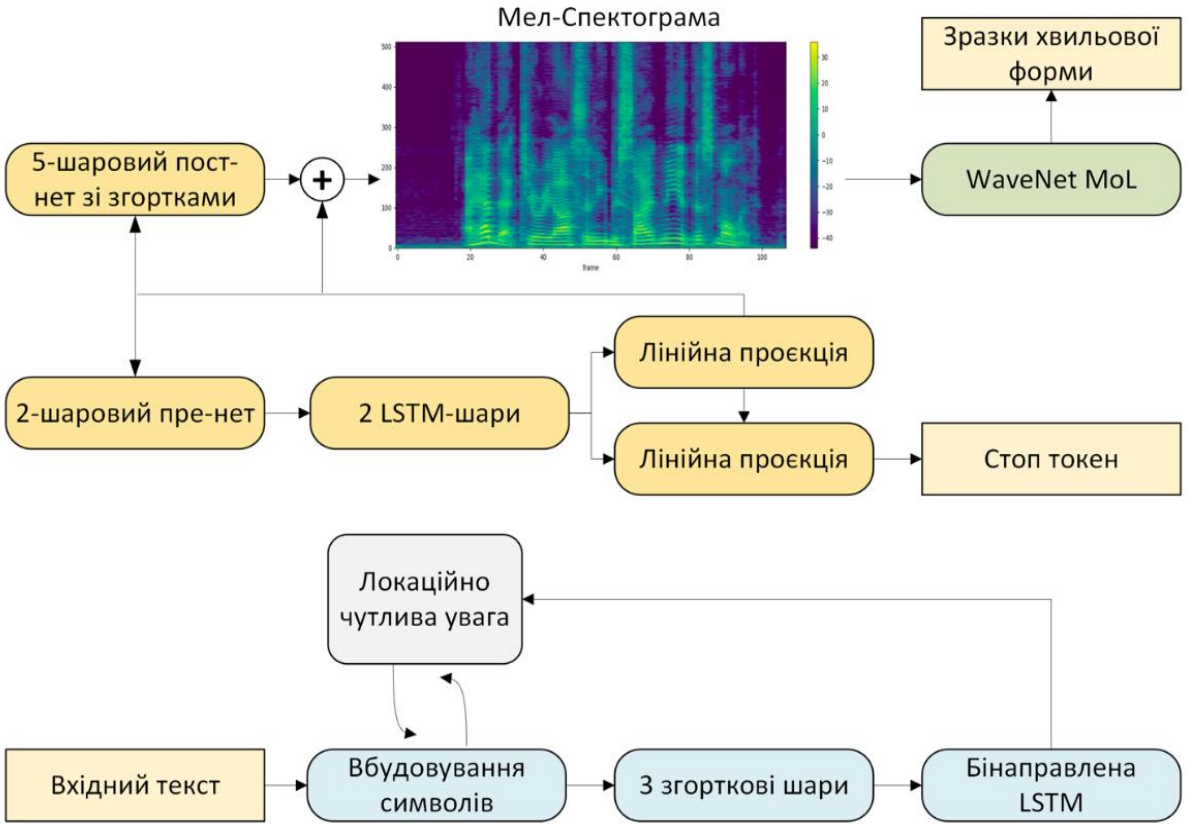

Рис. 3. Архітектурна модель Tacotron 2 [10]

Попри переваги також є і недоліки, такі як інтонація та обчислювана вартість. Незважаючи на високу якість, Tacotron 2 може іноді відтворювати мовлення з не зовсім природною інтонацією, тому якщо говорити в контексті атак — це суттєвий мінус. Також навіть після оптимізації, WaveNet залишається складним для обчислення, що може призводити до затримок.

Загалом, Tacotron 2 продовжує розвивати успіх технології WaveNet, показуючи, як добре нейронні мережі можуть синтезувати мовлення з просодією, при тому притримуючись рівня якості звуку WaveNet. За рахунок використання мел-спектрограм та WaveNet для генерування звукових хвиль, ця архітектура спрощує традиційні методи та показує, що системи можуть самостійно вивчати необхідні патерни, аби створювати природне мовлення. Цю систему можна навчати безпосередньо з даних, не покладаючись на складну інженерію функцій, і вона досягає якості звуку, яка близька до природної людської мови.

VITS

VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) — це модель наскрізного синтезу мовлення, що поєднує варіаційні автокодери (VAE) із GAN (генеративними змагальними мережами). VITS виділяється своєю здатністю виробляти різноманітний і виразний синтез голосу, вловлюючи емоційні тони та нюанси мови. Вона складається з постериорного кодера, декодера та умовного апріора, що дозволяє генерувати мовленнєві хвилі в залежності від вхідного тексту. Архітектура VITS наведена в табл. 3 (також див. рис. 4).

Таблиця 3

Архітектура VITS

Компонент	**Опис**	**Функція**
Умовний варіаційний автокодер (VAE)	Містить постеріорний енкодер, декодер та умовний апріор. Постеріорний енкодер перетворює лінійні спектрограми в латентний простір zzz, що описує основні характеристики аудіосигналу. Умовний апріор прогнозує розподіл zzz для відповідного тексту і разом із декодером відтворює аудіосигнал	Генерація латентних представлень аудіосигналу з тексту
Текстовий енкодер	Обробляє вхідні фонеми та перетворює їх у відповідні абстрактні репрезентації. Використовує нормалізуючі потоки для поліпшення апріорного розподілу і прогнозування аудіосигналів	Перетворення текстових послідовностей у векторні подання
Стохастичний предиктор тривалості	Дозволяє моделі генерувати мовлення з різними ритмами для одного і того ж тексту. Це підвищує природність та виразність синтезованого мовлення, генеруючи різні інтонації і паузи для одного тексту	Генерація варіативної тривалості та ритму мовлення
Декодування аудіосигналу	Латентні представлення з VAE декодуються стеками транспонованих згорткових шарів, схожих на HiFi-GAN, для точного та натурального перетворення спектрограм у звукові сигнали	Перетворення латентних векторів у реальний мовленнєвий сигнал

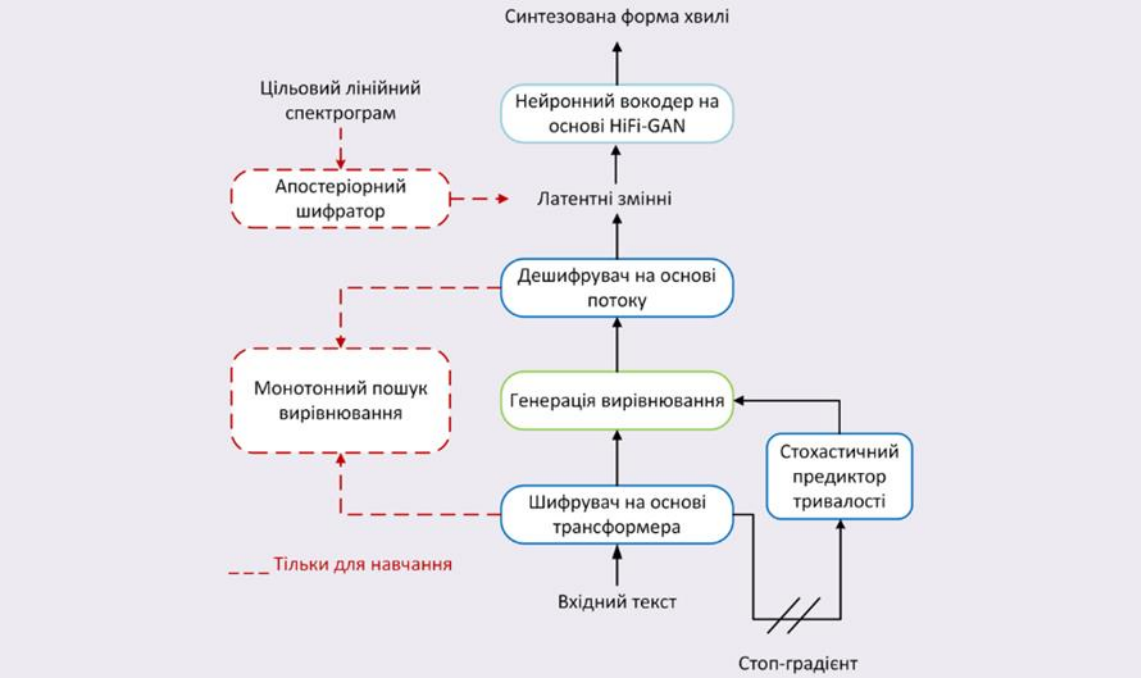

Рис. 4. Архітектурна модель VITS [12], [13]

Якщо говорити про Workflow даної моделі (рис. 5), можна побачити, що вхідне референційне аудіо (*Reference Audio*) та текст подаються у блоки, що містять енкодер, нормалізуючі потоки, і декодер. *Tone Extractor* обробляє вхідне аудіо та текст для вилучення важливих тональних характеристик, які потім обробляються у потоці, включаючи варіативні латентні простори для більшої гнучкості. На виході отримуємо синтезоване мовлення із вбудованим тоном і ритмом завдяки *Tone Converter* і додатковим модулям передбачення тривалості та текстовому кодуванню.

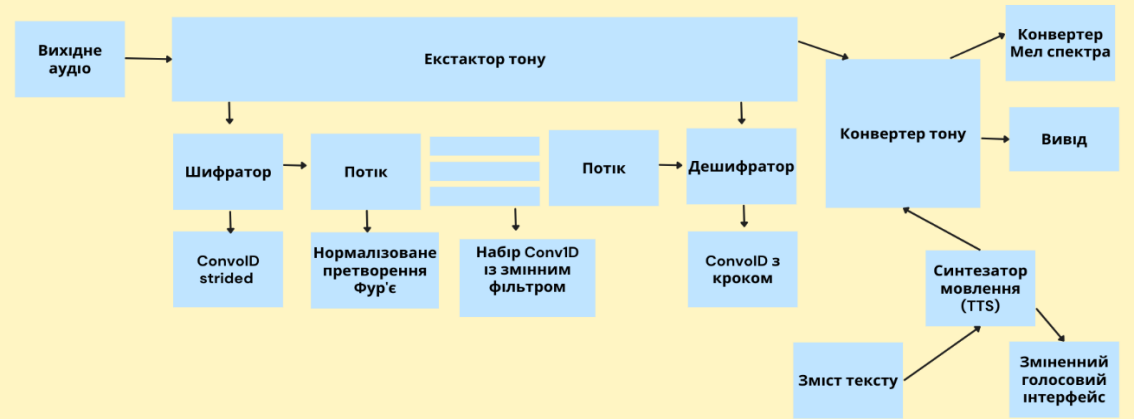

Рис. 5. Архітектура Workflow

Отож, в даної моделі є три основні переваги. Перша — це ще краща якість звуку. VITS йде ще далі, поєднучи кращі характеристики своїх попередників WaveNet та Tacotron 2 та використовуючи варіаційні автоенкодери. Також серед порівнюваних трьох моделей лише VITS може генерувати звуки не лише з тексту, але і з шумових сигналів.

VITS забезпечує швидку обробку даних під час **генерації звуку**, що зменшує вимоги до апаратних ресурсів. Тобто, коли модель вже навчена і використовується для створення звуку, вона працює швидше і з меншою кількістю ресурсів.

Однак, є і недоліки — на етапі **навчання** модель потребує багато ресурсів. Її архітектура є складною, включає багато компонентів і підходів, що ускладнює налаштування та оптимізацію. Щоб досягти найкращих результатів, VITS потребує значних обсягів даних і обчислювальних потужностей під час тренування. Тобто, модель вимагає багато ресурсів під час навчання, але працює швидко і ефективно після цього.

Тож розроблення моделі VITS можна вважати ще одним важливим кроком вперед у сфері технологій синтезу мовлення.

Порівняння ефективності клонування для атак

Ефективність клонування голосу для обходу біометричних систем значно зросла завдяки вдосконаленню технологій, таких як Tacotron 2, WaveNet, і VITS. Приклади атак демонструють, наскільки легко клонування голосу на основі ГНМ для обходу біометричних систем можуть бути використані зловмисниками.

Tacotron 2 та WaveNet здатні генерувати реалістичний голос, який звучить природно і включає інтонаційні особливості мовлення. Проте, обидві ці технології можуть мати певні затримки, що знижує їх ефективність у режимі реального часу. VITS, з іншого боку, використовує варіаційні автоенкодери для швидшого генерування голосу без значної втрати якості. Тобто для атак, де потрібна негайна автентифікація, це є більш привабливим варіантом і, ймовірно, зловмисних обере його, якщо не буде інших вагомих факторів, які здатні змінити тактику (рис. 6).

ISSN 2663 - 4023

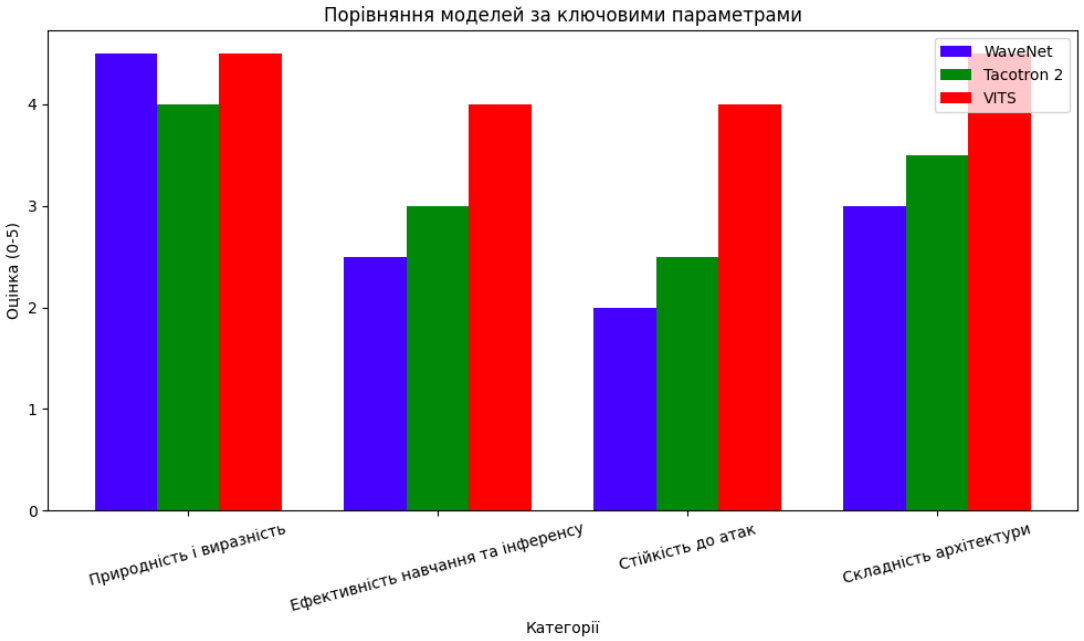

Рис. 6. Порівняння моделей WaveNet, Tacotron 2, VITS

Реальні випадки підтверджують небезпеку використання технологій клонування голосу. Хоча прямі докази використання конкретних моделей, таких як Tacotron 2, WaveNet або VITS, у наступних прикладах відсутні, та це може бути пов’язано з тим, що інформація про конкретні технології, які використовують зловмисники, рідко стає публічно доступною, оскільки це частина методів, що приховують їхні дії. Наприклад, в одному з інцидентів 2020 року шахраї використовували синтетичний голос, щоб імітувати керівника компанії з Німеччини і переконати фінансового директора в Великій Британії перевести $243,000 на їхній рахунок. Атака була здійснена за допомогою AI-генератора голосу, який зміг відтворити природну інтонацію, тембр і вимову керівника, що зробило підробку надзвичайно важко розпізнати [14]. Можна припустити, що було використано ГНМ, бо це схоже на методи, реалізовані в таких моделях, як Tacotron 2 або VITS, але точну модель не було розкрито.

Інший відомий випадок стосується використання технології, яку не розглянуто а даній статті, проте для загальної картини варто розглянути. Технологія ElevenLabs була використана для обходу системи голосової автентифікації Lloyds Bank. Дослідник зміг згенерувати голос, який успішно проходив перевірку банку, після деяких налаштувань синтетичного голосу, щоб краще передати природні ритми та інтонації мовлення. Це свідчить про те, що навіть відносно прості та доступні інструменти можуть бути використані для обходу систем, які вважаються надійними [15].

Можна зробити висновок, що успішність атак за допомогою клонування голосу значно залежить від якості тренувальних даних і можливостей алгоритмів адаптувати синтезовані голоси до різних сценаріїв. Саме хехнології типу VITS, завдяки своїй стохастичній природі, презентують себе краще у створенні природного голосу, що дозволяє успішно обійти системи навіть у режимі «zero-shot», тобто без значних попередніх даних про цільовий голос. Саме тому ці моделі особливо небезпечні в контексті атак на біометричні системи автентифікації, адже можна генерувати підробки з мінімальними вимогами до початкових даних.

Ці приклади демонструють, наскільки вразливими можуть бути сучасні системи голосової автентифікації перед загрозою клонування голосу, і підкреслюють необхідність впровадження додаткових заходів безпеки для зменшення ризику таких атак.

Технічна оцінка стійкості систем до атак клонування голосу

Польськими дослідниками було проаналізовано сучасні методи автентифікації мовця, де було підкреслено актуальні проблеми та обмеження цього підходу в умовах зростаючих атак клонування голосу та інших видів підробок [6]. Розробка стійких до клонування моделей автентифікації потребує вдосконалення методів, здатних протидіяти атакам з використанням нейронних мереж та технологій синтезу голосу.

Тому експеримент мав на меті розробити та протестувати модель надійної системи автентифікації мовця на основі біометричних характеристик голосу на платформі Deep Speaker. Розроблена модель повинна була коректно розпізнавати індивідуальні параметри голосу, зберігаючи високу точність навіть у випадку значного розширення бази даних.

Для побудови моделі використовувалися ГНМ (зокрема, згорткові та рекурентні нейронні мережі, CNN і RNN), які здатні аналізувати акустичні сигнали та виділяти специфічні патерни голосу, такі як частотні характеристики (аналіз спектральних коефіцієнтів голосового сигналу, що дає можливість виявляти унікальні частоти) та акустичні параметри (тембр, висота, швидкість мовлення та інші індивідуальні риси, що використовуються для ідентифікації мовця). Було використано великий набір аудіоданих, що включав записи мовців із різними тембрами, частотними та ритмічними параметрами. Розрахункові дані включали коефіцієнти мел-кепстральних характеристик (MFCC), які широко застосовуються у розпізнаванні мови для виділення частотних характеристик голосу мовця.

На етапі підготовки даних розраховувалися ключові характеристики кожного запису, такі як мел-кепстральні коефіцієнти (MFCC) та форманти (F1, F2, F3). Мел-кепстральні коефіцієнти (MFCC) розраховані шляхом сегментації сигналу на фрейми (10–20 мс), розрахунок частотних коефіцієнтів та побудови спектрограм для кожного мовця, а форманти (F1, F2, F3) базуються на аналізі основних частотних компонентів голосу мовця, зокрема першого, другого та третього форманту, які відображають унікальні акустичні характеристики мовлення кожної особи.

Дослідження також включає метод навчання з триплетними втратами для підвищення точності системи. Використання триплетів, що складаються з одного позитивного зразка, одного зразка-прив’язки та 99 негативних, дозволяє моделі ефективніше розрізняти різні зразки голосу, покращуючи надійність системи. Збільшення кількості епох для триплетних втрат покращує узагальнюючу здатність моделей, зберігаючи їх стійкість до атак.

Для оцінки подібності між голосовими зразками використовується косинусна подібність, яка визначає, наскільки близько два вектори (які представляють голосові характеристики) спрямовані один до одного. Формула для косинусної подібності виглядає так (1):

$$C_S = \frac{\text{A}\cdot\text{B}}{||\text{A}||\cdot||\text{B}||} = \frac{\sqrt{\sum_{i=1}^{n} A_i B_i}}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

де A і B — вектори, що представляють акустичні характеристики голосу.

Косинусна подібність коливається від -1 до 1, де 1 вказує на повну подібність, а -1 — на повну відмінність. Вона використовується для визначення, наскільки близько два зразки голосу можуть бути один до одного, і є важливим критерієм для оцінки якості автентифікації.

ISSN 2663 - 4023

Оцінка моделей виявила значні відмінності в точності та показниках EER (Equal Error Rate). Найнижча точність склала 96% з EER 5,6%, вказуючи на більший ризик помилкових спроб, тоді як найвища точність досягла 99,6% з EER 2,5%, що свідчить про високу стійкість до атак та точне розпізнавання автентичних користувачів (рис. 7).

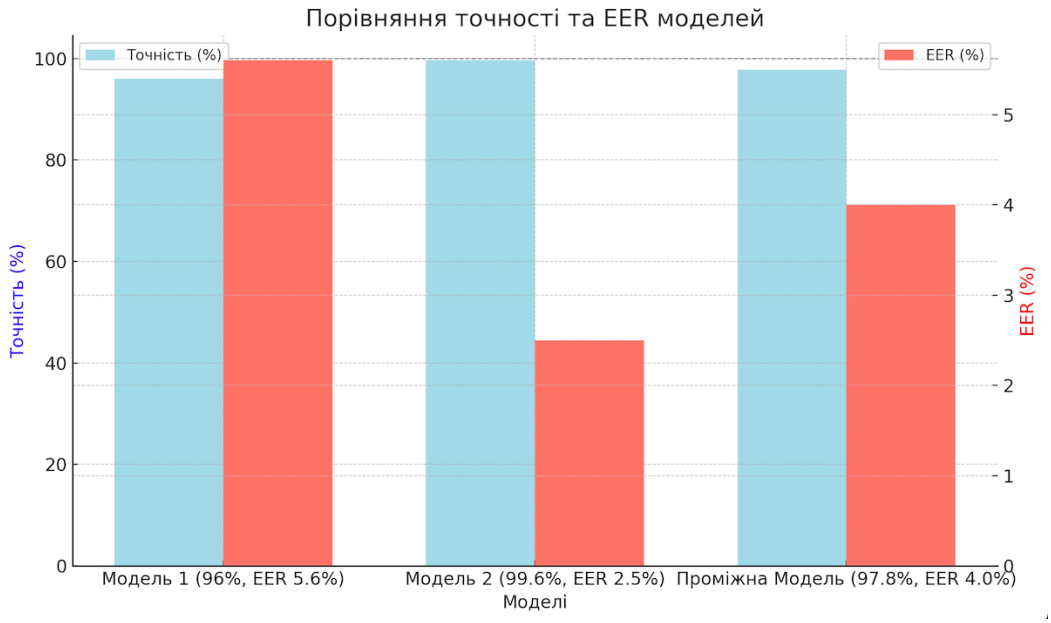

Рис. 7. Порівняння точності та показників EER

Рівна частота помилок (EER) є критичним показником для визначення балансу між рівнями хибного прийняття (FAR) і хибного відхилення (FRR). Формули для розрахунку FAR і FRR (2), (3):

$$FAR = \frac{FP}{(FP+TN)} \quad (2)$$

$$FAR = \frac{FN}{(FN+TP)} \quad (3)$$

де FP — помилкові позитивні результати, TN — справжні негативні результати, FN — помилкові негативні результати, TP — справжні позитивні результати.

EER є точкою, в якій FAR і FRR перетинаються, і її мета — забезпечити баланс між помилковими прийняттями та відхиленнями. Оптимізація системи часто намагається досягти якомога нижчого EER.

Паралельно з об'єктивними метриками, було проведено опитування слухачів для оцінки автентичності голосів. Метод парного порівняння показав, що деякі клоновані зразки важко відрізнити від оригіналів, підкреслюючи високу якість клонування. Середня похибка вказує на стабільність відповідей учасників, хоча якісні клоновані голоси залишалися складними для ідентифікації.

Отож, моделі з триплетним навчанням продемонстрували високу стійкість до атак клонування голосу, особливо при достатній кількості епох для триплетних втрат. Вибір моделей з низьким EER забезпечує високу точність і мінімальний ризик помилкових спроб. Однак суб'єктивні оцінки вказують на реалістичність деяких клонованих голосів, що вимагає подальшого вдосконалення біометричних систем для покращення їх стійкості. Система Deep Speaker показала свою здатність протистояти атакам голосового клонування, показуючи ефективність сучасних методів синтезу голосу у забезпеченні безпеки голосової автентифікації.

Складність розпізнання атак та аналіз методів захисту біометричних систем

Технології клонування голосу, особливо ті, що використовують ГНМ, значно вдосконалилися, тому це створює проблеми для біометричних систем та подальшого виявлення атак.

Адже за допомогою описаних вище WaveNet, Tacotron 2 і VITS, штучно створені голоси вкрай важко візуально та акустично відрізнити від реальних. Ці моделі здатні генерувати мову з надзвичайно високим рівнем природності, включаючи емоції, інтонації, тембри та навіть паузи.

Наприклад, осягнути рівень складності розпізнавання синтезованого та людського голосу можна на сайті веб-сайті GitHub [16] – [17], продемонстровано голос, який створений системою Tacotron 2 і який звучить настільки людським, що його майже неможливо розрізнити. Можна почути два жіночі голоси: один людський, а інший штучний інтелект, які вимовляють одне й те саме речення: *«She earned a doctorate in sociology at Columbia University»*.

Також варто згадати атаки «zero-shot», тобто можливість працювати з обмеженою або навіть відсутньою кількістю тренувальних даних. Це означає, що зловмисники можуть клонувати голоси людей, не маючи доступу до їхніх оригінальних записів, а системи автентифікації не можуть надійно відрізнити синтетичні голоси, засновані на зразках, які не були використані для навчання.

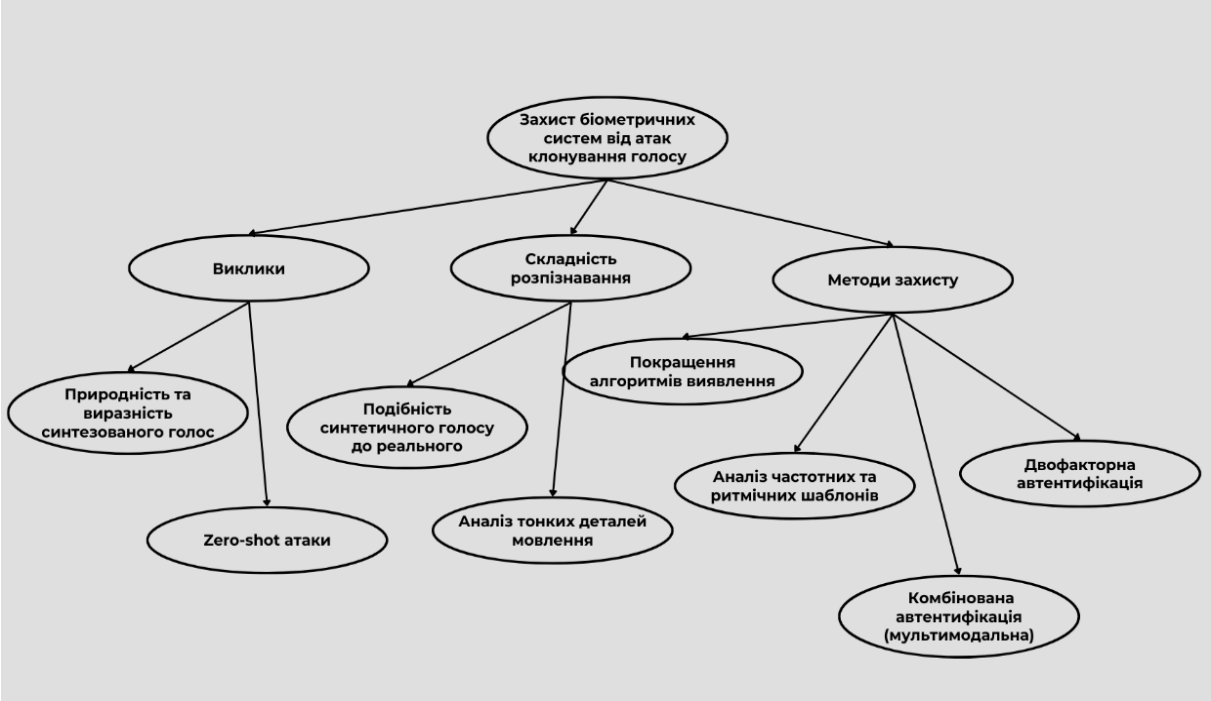

Рис. 8. Основні параметри захисту біометричних систем від атак на основі клонування голосу

Такий прогрес відкриває нові можливості для розвитку систем штучного інтелекту та голосових асистентів. Однак цей прорив також став серйозним інструментом для біометричних систем, оскільки зловмисники можуть використовувати синтетичні голоси для атак. Традиційні методи виявлення вже не можуть ефективно розрізняти справжні голоси та підроблені, багато систем не встигають адаптовуватися до нових методів атак, тому загроза безпеці таких систем неминуча.

З огляду на складність розпізнавання атак, необхідно постійно вдосконалювати алгоритми виявлення синтетичного мовлення та розвивати комбіновані методи автентифікації, які поєднують різні біометричні характеристики для підвищення надійності систем безпеки. Рекомендується активно досліджувати нові підходи та технології, щоб забезпечити належний рівень захисту від зловмисних атак.

Один із основних підходів — розробка алгоритмів для виявлення нехарактерних для людського мовлення частотних або ритмічних шаблонів, а також аналіз поведінкових факторів. Складність полягає в можливості технологій імітувати навіть найтонші деталі в мовленні, що ускладнює процес розпізнавання атак.

Важливо звернути увагу та використання ГНМ з точки зору захисту. Попри значні обчислювальні ресурси та велику кількість навчальних даних для досягнення високої точності — такі мережі можуть бути налаштовані для ідентифікації ледь помітних відмінностей між реальним і синтезованим голосом.

Також для підвищення безпеки рекомендується використовувати мультимодальні біометричні системи, наприклад, комбінувати голосову біометрію з іншими формами автентифікації, такими як розпізнавання обличчя, відбитки пальців або двофакторна автентифікація. Диверсифікація зменшує ймовірність успіху атаки, адже якщо навіть один з методів буде скомпрометовано, це не надасть доступу для входу.

В цілому, можна досягнути багаторівневого захисту, поєднуючи та вдосконалюючи усі методи.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У даній науковій роботі досліджено стійкість біометричних систем автентифікації до атак із застосуванням технології клонування голосу на основі глибинних нейронних мереж, а саме WaveNet, Tacotron 2 і VITS. WaveNet здатний створювати природний звук, проте обмеження в швидкості обробки можуть бути недоліком для реального використання. Tacotron 2 забезпечує високу якість синтезу з емоційними інтонаціями, але має певні проблеми з затримкою та вимогами до обчислювальних ресурсів. VITS, у свою чергу, поєднує в собі кращі якості своїх попередників, забезпечуючи швидкість та адаптивність, що в контексті атак на біометричні системи автентифікації є небезпечним.

Згідно досліджень, люди можуть легко заплутатися, вказавши на клоновані записи як на оригінальний зразок. У окремих випадках близько половини опитаних вказували на неправильну відповідь, плутаючи синтезований запис з оригінальним. Оцінюючи якість клонованих і справжніх записів після зміни частоти дискретизації, люди в більшості випадків вказували на подібну якість, а у деяких випадках якість клонованих записів навіть оцінювалася як «Добре», тоді як оригінальні — як «Задовільно». Водночас під час спроб атаки на розроблену біометричну систему майже всі спроби верифікації з клонованими зразками зазнали невдачі. Це підтверджує, що біометричні системи, засновані на ГНМ, здатні ефективно ідентифікувати синтезовані зразки, навіть коли люди мають проблеми з їх розрізненням.

Загалом, хоча попереду ще багато викликів, потенціал технологій клонування голосу на основі глибинних нейронних мереж безмежні. Еволюція досліджених моделей свідчить про значний прогрес у цій галузі. Проте, дивлячись у майбутнє, деякі сфери потребують подальших інновацій та розвитку. З огляду на це, є необхідність постійного вдосконалення алгоритмів виявлення синтетичного мовлення. Важливо розробляти нові методи, які не лише виявлятимуть характерні особливості синтетичних голосів, але й

комбінуватимуть різні біометричні параметри для підвищення надійності. Об'єднання голосової біометрії з іншими формами автентифікації, такими як розпізнавання обличчя або двофакторна автентифікація, може створити багаторівневий захист, зменшуючи ризик успішних атак.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Олешко, І. (2012). Порівняльний аналіз методів біометричної автентифікації на основі критерію відносної ентропії. *Вісник національного університету «Львівська політехніка»: Автоматика, вимірювання та керування, 741.*
2. Кустов, А., (2020). Спуфінг-атаки на біометричні системи автентифікації та методи протидії атакам. *Радіоелектроніка та молодь в XXI столітті: матеріали 24 Міжнародного молодіжного форуму, 5*, 76–77.
3. Кіщенко, М. І., & Пастушенко, М. С. (2021). Напрямки підвищення ефективності голосових систем автентифікації. *Сьома Міжнародна науково-технічна конференція «Проблеми електромагнітної сумісності перспективних безпроводових мереж зв'язку (ЕМС-2021)»*, 20–23.
4. Mohammadi, A., Sood, K., Nazari, A., & Thiruvady, D. (2024). *Securing Voice Authentication Applications Against Targeted Data Poisoning*. https://doi.org/10.48550/arXiv.2406.17277
5. *Approaches to Address AI-enabled Voice Cloning*. (2024). https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/04/approaches-address-ai-enabled-voice-cloning
6. Milewski, K., Zaporowski, S., & Czyżewski, A. (2023). Comparison of the Ability of Neural Network Model and Humans to Detect a Cloned Voice. *Electronics, 12(21)*. https://doi.org/10.3390/electronics12214458
7. Максименко, О. А. (2019). *Дипломна робота на здобуття ступеня бакалавра: «Генерація цільового голосу людини з використанням нейронних мереж»*. Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».
8. Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *Wavenet: a generative model for raw audio*. https://doi.org/10.48550/arXiv.1609.03499
9. Victor, A. O., & Ali, M., I. (2024). Enhancing Time Series Data Predictions: A Survey of Augmentation Techniques and Model Performance. *ACSW'24: Proceedings of the 2024 Australasian Computer Science Week*, 1–13. https://doi.org/10.1145/3641142.364114
10. Shen, J., et al. (2017). *Natural tts synthesis by conditioning wavenet on mel spectrogram predictions*. https://doi.org/10.48550/arXiv.1712.05884
11. Chapuzet, A. (n. d.). *Speech Synthesis (TTS), How to Use It and Why Is It So Important?* https://vivoka.com/how-to-speech-synthesis-tts
12. Verma, U., & Padmanaban, R. (2024). Speech Cloning: Text-To-Speech Using VITS. *Engineering and Technology Journal, 9(5)*. https://doi.org/10.47191/etj/v9i05.10
13. Kim, J., Kong, J., & Son, J. (2021). *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech*. https://doi.org/10.48550/arXiv.2106.06103
14. Malyshev, A. (2023). *Voice Cloning: A Blessing or a Curse for the Voice Banking Industry?* https://www.finextra.com/blogposting/23813/voice-cloning-a-blessing-or-a-curse-for-the-voice-banking-industry
15. Cox, J. (2023). *How I Broke Into a Bank Account With an AI-Generated Voice*. https://www.vice.com/en/article/how-i-broke-into-a-bank-account-with-an-ai-generated-voice/
16. *Audio samples from "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions"*. (n. d.). https://google.github.io/tacotron/publications/tacotron2/index.html
17. Гулак, Г. М., Жильцов, О. Б., Киричок, Р. В., Коршун, Н. В., & Складанний, П. М. (2024). *Інформаційна та кібернетична безпека підприємства.* Підручник. Львів : Видавець Марченко Т. В.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (26), 2024
ISSN 2663 - 4023

Tatiana Savkova
Student of of the Department of Information Protection
Lviv Polytechnic National University, Lviv, Ukraine
ORCID ID: 0009-0009-1112-9508
tetiana.savkova.kb.2021@lpnu.ua

Ivan Opirskyy
Doctor of Technical Sciences, Professor, Head of the Department of Information Protection
Lviv Polytechnic National University, Lviv, Ukraine
ORCID ID: 0000-0002-8461-8996
ivan.r.opirskyi@lpnu.ua

Dmytro Sabodashko
PhD, Senior Lecturer of the Department of Information Protection
Lviv Polytechnic National University, Lviv, Ukraine
ORCID ID: 0000-0003-1675-0976
dmytro.v.sabodashko@lpnu.ua

STUDYING THE RESISTANCE OF BIOMETRIC AUTHENTICATION SYSTEMS TO ATTACKS USING VOICE CLONING TECHNOLOGY BASED ON DEEP NEURAL NETWORKS

Abstract. With the development of voice synthesis technologies based on deep neural networks, the security threats to biometric authentication systems that use voice recognition have increased. These systems, which were considered reliable, can be easily compromised by fake voices created using advanced models such as WaveNet, Tacotron 2, and other modern algorithms. In today's cybersecurity environment, such attacks jeopardize the confidentiality of personal data, which necessitates the improvement of protection methods. The purpose of this article is to study the resilience of biometric authentication systems to attacks using voice cloning technology, to analyze the effectiveness of modern synthesis methods for circumventing such systems, and to provide a comparative overview of various approaches to protect voice biometric data. The article discusses technologies that allow for the creation of accurate and realistic synthetic voices, as well as methods for detecting and protecting against fake signals. The article also analyzes the current vulnerabilities of voice systems and suggests strategies to increase resistance to such attacks, providing users with greater security and privacy.

Keywords: voice cloning; biometric authentication systems; deep neural networks; WaveNet; Tacotron 2; security; voice synthesis.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Oleshko, I. (2012). Comparative analysis of biometric authentication methods based on the relative entropy criterion. *Bulletin of Lviv Polytechnic National University: Automation, Measurement and Control, 741.*
2. Kustov, A., (2020). Spoofing attacks on biometric authentication systems and methods of countering attacks. *Radio electronics and youth in the XXI century: materials of the 24th International Youth Forum, 5*, 76–77.
3. Kishchenko, M. I., & Pastushenko, M. S. (2021). Directions for improving the efficiency of voice authentication systems. *Seventh International Scientific and Technical Conference "Problems of electromagnetic compatibility of advanced wireless communication networks (EMC-2021)",* 20–23.
4. Mohammadi, A., Sood, K., Nazari, A., & Thiruvady, D. (2024). *Securing Voice Authentication Applications Against Targeted Data Poisoning*. https://doi.org/10.48550/arXiv.2406.17277
5. *Approaches to Address AI-enabled Voice Cloning.* (2024). https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/04/approaches-address-ai-enabled-voice-cloning
6. Milewski, K., Zaporowski, S., & Czyżewski, A. (2023). Comparison of the Ability of Neural Network Model and Humans to Detect a Cloned Voice. *Electronics, 12(21)*. https://doi.org/10.3390/electronics12214458

7. Maksymenko, O. A. (2019). *Bachelor's thesis: "Generation of the target human voice using neural networks".* National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".
8. Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *Wavenet: a generative model for raw audio*. https://doi.org/10.48550/arXiv.1609.03499
9. Victor, A. O., & Ali, M., I. (2024). Enhancing Time Series Data Predictions: A Survey of Augmentation Techniques and Model Performance. *ACSW'24: Proceedings of the 2024 Australasian Computer Science Week*, 1–13. https://doi.org/10.1145/3641142.364114
10. Shen, J., et al. (2017). *Natural tts synthesis by conditioning wavenet on mel spectrogram predictions*. https://doi.org/10.48550/arXiv.1712.05884
11. Chapuzet, A. (n. d.). *Speech Synthesis (TTS), How to Use It and Why Is It So Important?* https://vivoka.com/how-to-speech-synthesis-tts
12. Verma, U., & Padmanaban, R. (2024). Speech Cloning: Text-To-Speech Using VITS. *Engineering and Technology Journal, 9(5).* https://doi.org/10.47191/etj/v9i05.10
13. Kim, J., Kong, J., & Son, J. (2021). *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech.* https://doi.org/10.48550/arXiv.2106.06103
14. Malyshev, A. (2023). *Voice Cloning: A Blessing or a Curse for the Voice Banking Industry?* https://www.finextra.com/blogposting/23813/voice-cloning-a-blessing-or-a-curse-for-the-voice-banking-industry
15. Cox, J. (2023). *How I Broke Into a Bank Account With an AI-Generated Voice*. https://www.vice.com/en/article/how-i-broke-into-a-bank-account-with-an-ai-generated-voice/
16. *Audio samples from "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions".* (n. d.). https://google.github.io/tacotron/publications/tacotron2/index.html
17. Hulak, H. M., Zhiltsov, O. B., Kyrychok, R. V., Korshun, N. V., & Skladannyi, P. M. (2024). *Information and cyber security of the enterprise.* Textbook. Lviv: Publisher Marchenko T. V.