Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Метод вбудови цифрових водяних знаків в аудіо-файли за допомогою машинного навчання

(тема)

Виконав:

студент II курсу, групи КСМм-22-1

Гунько М. О.
(прізвище, ініціали)

Спеціальність 123 «Комп’ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Комп’ютерні системи та мережі
(повна назва освітньої програми)

Керівник: доц. Мартовицький В.О.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ЕОМ _____ Коваленко А.А.
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп’ютерна інженерія»
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Комп’ютерні системи та мережі
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
“____” _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Гуньку Максиму Олексійовичу
(прізвище, ім’я, по батькові)

1. Тема роботи Метод вбудови цифрових водяних знаків в аудіо-файли за допомогою машинного навчання

затверджена наказом по університету від “ 06 ” листопада 2023 р. № 1298 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 15 січня 2024 р.

3. Вхідні дані до роботи Набір даних для навчання

4. Перелік питань, що потрібно опрацювати у роботі

Аналіз методів стеганографії в аудиофалах

Аналіз нейромережних архітектур

Розробка методу вбудови цифрових водяних знаків в аудіо-файли за допомогою машинного навчання

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів)

Слайд-презентація – 13 слайдів

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Огляд автентифікації аудіо-даних	07.11.23 – 05.11.23	
2	Огляд механізмів захисту цілісності	06.11.23 – 12.11.23	
	аудіоданих		
3	Огляд методів навчання	13.11.23 – 19.11.23	
4	Розробка методу захисту аудиоданих	20.11.23 – 03.12.23	
	За допомогою нейромережі		
5	Проведення експериментів	04.12.23 – 10.12.23	
6	Оформлення матеріалів кваліфікаційної роботи	11.12.23 – 29.12.23	
7	Подання кваліфікаційної роботи керівникові та	30.12.23 – 04.01.24	
	її попередній захист		
8	Подання кваліфікаційної роботи на	05.01.24 – 10.01.24	

Дата видачі завдання 06 листопада 2023 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Мартовицький В. О.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 52 с., 21 рис., 4 табл., 1 дод., 19 джерел.

СТЕГАНОГРАФІЯ, АУДИОФАЙЛИ, ФУРЬЄ, ЦВЗ, НЕЙРОМЕРЕЖА, МАШИННЕ НАВЧАННЯ.

У даній кваліфікаційній роботі досліджується метод вбудови цифрових водяних знаків в аудіо-файли з використанням технік машинного навчання. Робота вивчає різноманітні методи та алгоритми машинного навчання, які можуть бути використані для ефективної вбудови водяних знаків в аудіо-файли з метою захисту авторських прав та автентифікації контенту. Особлива увага приділяється вибору підходів машинного навчання, які найбільш ефективно враховують специфіку аудіо-даних, таких як спектрограми, характеристики звукових сигналів тощо. Досліджуються виклики та обмеження, що виникають у процесі вбудови цифрових водяних знаків в аудіо-файли, а також розглядаються можливі застосування та переваги використання таких методів для різних сфер, включаючи музичну індустрію, аудіо-відео контент та інші області. Результати дослідження можуть мати важливе значення для розвитку нових методів захисту авторських прав та забезпечення інтегритету аудіо-контенту у цифровому середовищі.

ABSTRACT

Master’s thesis: 52 pages, 21 figures, 4 tables, 1 appendices, 19 sources.

STEGANOGRAPHY, AUDIO FILES, FOURIER TRANSFORM, CWM, NEURAL NETWORK, MACHINE LEARNING.

This qualification work investigates a method of embedding digital watermarks in audio files using machine learning techniques. The work explores various machine learning methods and algorithms that can be used to effectively embed watermarks in audio files for copyright protection and content authentication purposes. Particular attention is paid to the selection of machine learning approaches that most effectively take into account the specifics of audio data, such as spectrograms, characteristics of audio signals, etc. The challenges and limitations that arise in the process of embedding digital watermarks in audio files are investigated, and the possible applications and benefits of using such methods for various fields, including the music industry, audio and video content, and other areas are considered. The results of the study may be important for the development of new methods of copyright protection and ensuring the integration of audio content in the digital environment.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

CCTV – відеоспостереження (англ., Closed Circuit Television)

BWF – це розширення імені аудіофайлів, збережених у цьому форматі (англ., Broadcast Wave Format)

rMAC – код автентифікації повідомлення (англ., Message Authentication Code)

RIPEMD – алгоритм безпечного (англ., RACE Integrity Primitives Evaluation Message Digest)

HMAC – ключовий код автентифікації повідомлень (англ., Hash-based Message Authentication Code)

ВСТУП

Піратство ліцензійних та захищених авторським правом аудіо- та музичних творів є проблемою протягом десятиліть. Оминаючи дозволені канали розповсюдження, контент поширюється без виплати винагороди авторам та правовласникам. Щоб зменшити використання цього незаконно отриманого контенту, необхідно мати можливість довести, що він не походить із законних джерел. Одним з інструментів для цього є позначення контенту певним повідомленням або метаданими перед розповсюдженням. При виявленні контенту, який підозрюється в незаконному поширенні, повідомлення може бути використане для відстеження його походження. Цей метод називається нанесенням водяних знаків.

Водяні знаки – це коли сам звук модифікується для передачі повідомлення. Це створює багато проблем, оскільки зміни в аудіо не повинні бути непомітними для слухача, але при цьому їх можна виявити при витягуванні повідомлення. Традиційні підходи до нанесення водяних знаків чутливі до шуму. Якщо закодоване аудіо змінено або піддано впливу шуму, витягти водяний знак буде неможливо. Коли власник контенту хоче виявити та ідентифікувати водяний знак на носії, йому потрібен доступ до цифрового аудіофайлу. Якщо використовуваний метод нанесення водяних знаків є стійким до шуму, що вноситься при відтворенні аудіо в ефірі, водяний знак можна вилучити, записавши відтворене аудіо.

Новітні методи нанесення водяних знаків на зображення використовують машинне навчання для вбудовування повідомлення в зображення. За допомогою цих методів можна зробити водяні знаки більш стійкими до різних типів шуму. Якби ці методи можна було адаптувати до аудіо, то можна було б створювати стійкі водяні знаки і в аудіо області.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Мета дослідження

Метою цієї дипломної роботи є розробка методу надійного нанесення водяних знаків на аудіо з використанням глибинного навчання. Точніше, надійного в тому сенсі, що водяний знак може бути вилучений після відтворення та запису аудіо в ефірі. Бездротовий зв'язок означає, що аудіо відтворюється з динаміка, а потім записується мікрофоном десь в іншому місці в кімнаті. Кодування повинно бути зроблено в чутних частотах, через обмеження в апаратному забезпеченні споживчого класу. Завдання включає побудову конвеєра кодера та декодера, які навчаються спільно з використанням імітованого шуму. Завдання, яке необхідно вирішити, включає розробку кодера і декодера та їх спільне навчання за допомогою диференційованої імітації шуму в ефірному каналі. Загальний вигляд такої системи показано на рисунку 1.1.

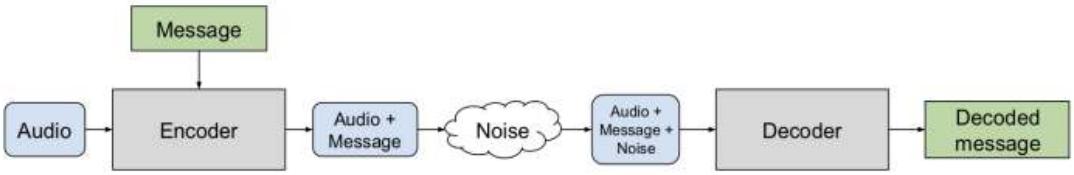

Рисунок 1.1 – Огляд системи для нанесення водяних знаків

1.2 Обмеження дослідження

Припускається, що закодоване аудіо не було змінено перед відтворенням. Це також включає припущення, що аудіо обмежується одним каналом і не стискається. Основна увага в роботі приділяється шумам, що

вносяться ефірним каналом, і тому ми обмежили сферу застосування роботи тільки цим.

Область, тісно пов'язана з нанесенням водяних знаків – це стеганографія. Метою стеганографії є вбудовування повідомлення в носій з додатковою метою зробити збурення носія невидимими для третьої сторони, наприклад, за допомогою статистичного аналізу закодованого файлу. У цій роботі ми не розглядаємо цей фактор і зосереджуємося виключно на тому, щоб зробити зміни непомітними при прослуховуванні.

Через накладні витрати на якісне дослідження якості звуку, якість звуку буде вимірюватися лише кількісно за допомогою відповідних метрик.

1.3 Достовірність цифрових аудіо-даних

Як і фотографія, історія звукозапису сягає 19 століття, а саме винаходу фонографа де Мартенвілем у 1860 році або фонографа Едісоном у 1877 році. Відтоді було створено незліченну кількість аналогових і цифрових аудіоносіїв. І в певних сценаріях відповідний зміст насправді міститься у фрагментах аудіоінформації, а не в зображеннях (у русі). Очевидно, що це стосується чистого аудіоконтенту, наприклад, записів голосу / передач або радіопередач. Але це також стосується звукових доріжок, що містяться у різного роду "відео" даних.

Вставка, склеювання, видалення, приглушення або обрізка аудіоматеріалів може бути легко застосована і може означати значну зміну того, що чує і, зрештою, розуміє користувач-людина. Наприклад, у голосовому записі навіть додана, змінена або видалена фраза короткої тривалості (наприклад, слова "так" або "ні") або префіксальний склад (наприклад, слова "згоден" або "не згоден") може повністю змінити його зміст. Навіть гірше, повні аудіозаписи можуть бути створені з нуля, що може приховати підроблене походження певного аудіодоказу. Як і у випадку з цифровими зображеннями, таку підробку можна легко застосувати і

приховати від пересічного слухача за допомогою програмного забезпечення для редагування аудіо та інших заходів: Відповідно, чути - не означає вірити!

Таким чином, достовірність і походження цих медіа з точки зору перевірки їхнього справжнього і неспотвореного стану є життєво важливими. Прикладами такого важливого контенту є історичні архівні матеріали, що зберігають культурну спадщину, та цифрові докази, про які йтиметься далі.

Чутливий аудіоконтент, який варто захищати, міститься в історичних архівних записах, що зберігають культурну спадщину. Нижче наведені найбільш релевантні класи і кілька мотивуючих прикладів такого контенту:

- історичні промови: Приклади відомих цитат з різних століть, які сприяють формуванню культурної пам'яті, є частиною загального знання. Записи таких подій доступні з кінця 19-го століття. Забезпечення їхньої доступності є природним мандатом державних архівів та подібних культурних інституцій;

- усноісторичні інтерв'ю: Крім того, державні архіви надають мільйони свідчень сучасних свідків з різних епох і культурних кіл. На відміну від попереднього прикладу з відомими публічними виступами, ці свідчення зберігають пам'ять і досвід незліченної кількості "звичайних" людей;

- новини: Новини про поточні події та справи в суспільстві також роблять свій внесок у сучасну історію. Те саме стосується зібраного новинного контенту, навіть якщо він не виходить в ефір, а залишається в інформаційних агентствах. Тут перевірка автентичності та цілісності джерел (запитів, репортерів, інформаторів, фотографів, знімальних груп тощо) завжди була головним обов'язком журналістів і письменників;

- користувацький контент: Будь-який вид створеного користувачами контенту, який циркулює в Інтернеті або в приватних колах, виражає інтереси та занепокоєння суспільства (і може бути, в свою чергу, знову опублікований в "Новинах");

- радіо/телепрограми: Навіть тривіальні повсякденні теле- і радіопрограми і навіть комерційна реклама в теле- і радіоефірі відображають культурне розмаїття сучасної цивілізації;

- музика: Серйозна музика та поп-музика є важливими елементами культурної спадщини також є важливими елементами культурної спадщини. Музика різних жанрів, створена/виконана різними виконавцями, композиторами чи диригентами. диригентами.

Рисунок 1.2 – Приклади культурної спадщини

Усі типи аудіоконтенту, згадані вище, доступні для громадськості у вигляді аналогових або цифрових/оцифрованих записів з архівів, музеїв чи інших культурних організацій, або є комерційними продуктами, або циркулюють в Інтернеті.

Інший важливий клас значущого аудіоконтенту – це докази, що містяться в цифровому аудіо. Приклади наведені нижче.

Допити в поліції та судові засідання: Свідчення свідка чи потерпілого або зізнання підозрюваного під час поліцейських допитів, судових засідань (у тому числі відеозаписів) часто записуються за допомогою електронного відеообладнання або простих диктофонів. У багатьох країнах (але не в Німеччині) це робиться в повному обсязі.

Законне перехоплення / електронне прослуховування: Підслуховування підозрюваних у скоєнні злочину здійснюється правоохоронними органами в ході кримінального розслідування. Тут аудіоспостереження за допомогою

електронних жучків на місці злочину або прослуховування телефонного зв'язку чи комп'ютера підозрюваного в ході законного (а також стратегічного) перехоплення є поширеним заходом для збору доказів.

Військові бойові операції: під час бойових дій бойові підрозділи та їхнє спорядження часто оснащені пристроями для передачі та запису відео в реальному часі. Голоси, шум навколишнього середовища (наприклад, постріли) або радіопереговори можуть бути використані для розслідування інцидентів або навіть воєнних злочинів.

Поліцейські операції в громадських місцях: часто поліцейські операції на великих публічних заходах із залученням великого скупчення людей (наприклад, демонстраціях, спортивних заходах) документуються поліцейськими нарядами на місці за допомогою відеокамер, з «боді-камерами», які носять на тілі чи приладах. кулачки в поліцейських патрульних автомобілях. Також протестувальники іноді знімають дії поліції на камери мобільних телефонів тощо. У разі протиправних дій будь-якої сторони ці кадри можуть бути опубліковані або передані до кримінальної чи цивільної справи.

Відеореєстратори: бортові "відеореєстратори" в автомобілях приватних користувачів, схоже, стають дедалі популярнішими в деяких країнах. Знову ж таки, звукова доріжка в таких відеозаписах, створених користувачами, може бути використана для обґрунтування цивільних позовів у разі дорожньо-транспортних пригод.

Відеоспостереження/CCTV: нарешті, все більшого поширення набувають системи відеоспостереження за громадськими місцями та приватними приміщеннями. За певних правових умов такі системи відеоспостереження (CCTV) також фіксують звук, який може бути використаний як доказ у кримінальних або цивільних справах.

Телефонний зв'язок: спілкування бізнесу з клієнтами по телефону вже десятиліттями є звичним явищем у сфері роздрібної торгівлі або домашнього банкінгу. Іноді колл-центри роблять запис телефонної розмови, щоб надати

докази в разі скарг клієнтів. Іншими прикладами таких важливих телефонних розмов є записи екстрених дзвінків до служб порятунку або поліції.

Цивільна авіація та судноплавство: Комерційні літаки та судна повинні бути обладнані записуючими пристроями (наприклад, диктофон в кабіні / бортовий самописець). У разі аварії або катастрофи, останні хвилини розмови екіпажу, оголошення для пасажирів та інші звуки/шуми повинні бути доступні для подальшого розслідування. Це також стосується радіопереговорів, записаних на місці розташування повітряних/морських диспетчерів.

Рисунок 1.3 – Приклади цифрових доказів

Наведені раніше приклади цифрових доказів об'єднує те, що життєво важливим є повний ланцюжок доказів. Як наслідок, запропоновані механізми захисту повинні бути інтегровані в (надійний) записуючий пристрій, наприклад, диктофон, камеру або в центр моніторингу.

Приклади аудіо-носіїв, що відповідають культурній спадщині та цифровим доказам, візуалізовано на рисунках 1.3 та 1.4. У світлі цих прикладів поняття достовірності аудіоданих буде більш детально розглянуто далі.

2 ОГЛЯД ВИКОРИСТАНИХ ПІДХОДІВ

2.1 Цифрові водяні знаки

Водяний знак – це повідомлення або метадані, вбудовані в медіа, такі як зображення або аудіо. Водяний знак вбудовується таким чином, щоб не впливати на сприйняття якості носія.

Повідомлення – це текстовий рядок, який вбудовується в зображення або аудіозапис, що також називається носієм. Типовий випадок використання водяних знаків складається з двох кроків. На першому етапі, кодуванні, повідомлення ховається в носії і виконується перед розповсюдженням аудіо або зображення. На другому етапі, декодуванні, правовласник перевіряє вже розповсюджений аудіофайл на наявність закодованого водяного знаку.

Нанесення водяних знаків не є новою ідеєю і вже давно розглядається як проблема чистої обробки сигналів. Класичні підходи включають в себе більш-менш кустарні рішення для кодування прихованої інформації в графічних та аудіо файлах. Ці підходи включають LSB-кодування [17], фазове кодування [2] та водяні знаки з розширеним спектром [5]. Лін та Абдулла [12] представляють огляд різних традиційних підходів до нанесення водяних знаків.

В останні роки з'явився ще один підхід до нанесення водяних знаків, який використовує машинне навчання для вивчення представлення інформації в засобах масової інформації. Балуджа [1] пропонує метод стеганографії в зображеннях, де передане повідомлення є іншим зображенням. Метод дозволяє передавати з втратами дуже великі повідомлення, хоча зміни носія в деяких випадках не є непомітними. Метод демонструє потенціал і вражаючі можливості підходів машинного навчання до нанесення водяних знаків і стеганографії.

Інший метод для стеганографії та нанесення водяних знаків на зображення запропонований Чжу та ін. в [18]. Вони допускають спотворення носія шумом після кодування, змушуючи вивчене представлення закодованого повідомлення бути стійким до внесеного шуму. Таким чином їм вдається зробити водяний знак стійким до стиснення JPEG з втратами.

Тема, тісно пов'язана з нанесенням водяних знаків, – це приклади зловмисників, область, яка часто досліджується в останні роки. Прикладом протидії є деякі дані (наприклад, зображення або аудіокліп), які були непомітно змінені, щоб обдурити вторинну систему (наприклад, класифікатор зображень або систему автоматичного розпізнавання мовлення). Подібність до нанесення водяних знаків вражає, де модифіковані дані є носієм, а вторинна система – декодером. Замість того, щоб обдурити вторинну систему, людина використовує модифікацію носія для розміщення водяного знаку. Одна з головних відмінностей полягає в тому, що в системі водяних знаків ми навчаємо вторинну систему разом з кодувальником. Цю область спочатку вивчали Сегеді та ін. в [16], а потім Гудфеллоу та ін. в [6]. Куракін та ін. [11] показують, що ці властивості нейронних мереж високої розмірності є стійкими до шуму.

Коли справа доходить до звукових сигналів, більшість дослідницьких методів намагаються обдурити автоматичні системи розпізнавання мовлення. Карліні та Вагнер [3] пропонують метод, який може успішно обдурити автоматичну систему розпізнавання мови, але зразки не залишаються змагальними при відтворенні в ефірі. Цинь та ін. [14] продовжують досліджувати цю задачу з метою зробити зразок змагальним в ефірі. Для цього вони імітують акустику кімнати під час тренування і використовують психоакустичну модель якості звуку. Вони показують, що їхні збурення зразків непомітні для людини, але вони не є повністю стійкими при відтворенні в ефірі.

2.2 Необхідні умови

2.2.1 Швидке перетворення Фур'є

Як має бути представлене аудіо під час обробки? Сире аудіо семплюється з високою частотою, зазвичай 44100 Гц для музики. Це призводить до дуже високої розмірності даних і залежностей між відліками, що простягаються далеко один від одного.

Альтернативою для представлення аудіо є використання короткого часового перетворення Фур'є (Short Time Fourier Transform, STFT). Виділяються сегменти хвильового файлу, і для кожного сегмента обчислюється відповідне перетворення Фур'є. Сегменти об'єднуються у двовимірне представлення звуку з часом на одній осі та частотами на іншій осі. Комплексне представлення часто виражається як амплітуда і фаза, де амплітуда також називається спектрограмою звуку. Значення на спектрограмі представляють енергію сигналу у відповідному частотному діапазоні, в кожному сегменті. Розмір сегментів задається розміром вікна, а відстань між початком кожного сегмента - довжиною стрибка. Зазвичай довжина стрибка менша за розмір вікна, так що вікна перекривають одне одного. На рисунку 2.1 відрізки відбираються з 50% перекриттям.

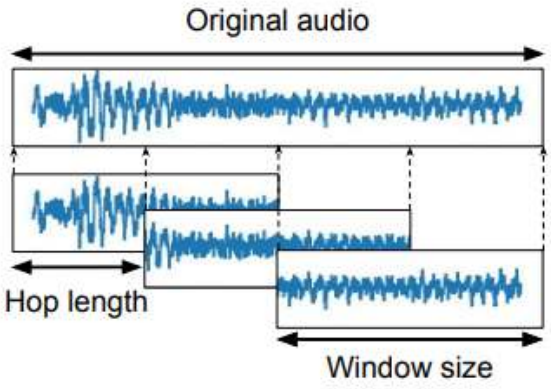

Рисунок 2.1 – Сегменти відбираються з форми сигналу з 50% перекриттям

З STFT можна без втрат відновити форму сигналу за допомогою зворотного STFT. Це дозволяє обробляти аудіо в STFT, використовуючи такі переваги, як спектральне представлення та зменшення часових залежностей, а потім реконструювати його до форми, придатної для відтворення. Рисунок 2.2 ілюструє такий конвеєр.

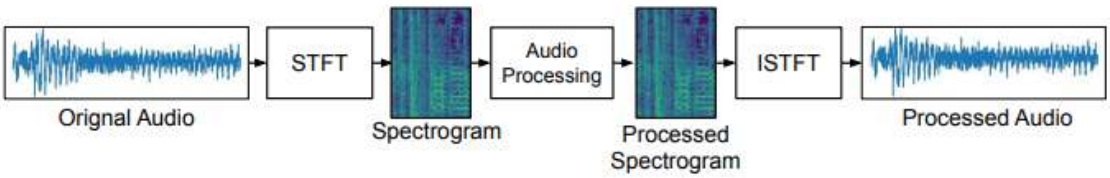

Рисунок 2.2 – Конвеєр для обробки звуку за допомогою STFT

2.2.2 Архітектура нейронних мереж

В обох методах Баджули [1] і Чжу та ін. [18] енкодери натхненні автокодерами. Автокодер – це мережева архітектура, де мережа повинна мінімізувати різницю між входом і виходом мережі, як правило, за певних обмежень, таких як зменшення розмірності або додавання шуму. У запропонованих методах нанесення водяних знаків обмеження полягає в тому, що закодований водяний знак повинен бути відновлюваним у декодері. Більше інформації про автокодери можна знайти в [7].

Архітектура U-net є прикладом автокодера, вперше представленого Роннебергером та ін. [15] як метод розділення джерел на медичних зображеннях. Модель спрямована на збереження низькорівневих деталей шляхом додавання пропускних зв'язків між низхідною та висхідною дискретизацією в автокодері. Зменшення просторової розмірності досягається за допомогою декількох шарів послідовної згортки, а збільшення – за допомогою відповідних шарів транспонованої згортки. Приклад архітектури U-мережі показано на рисунку 2.3.

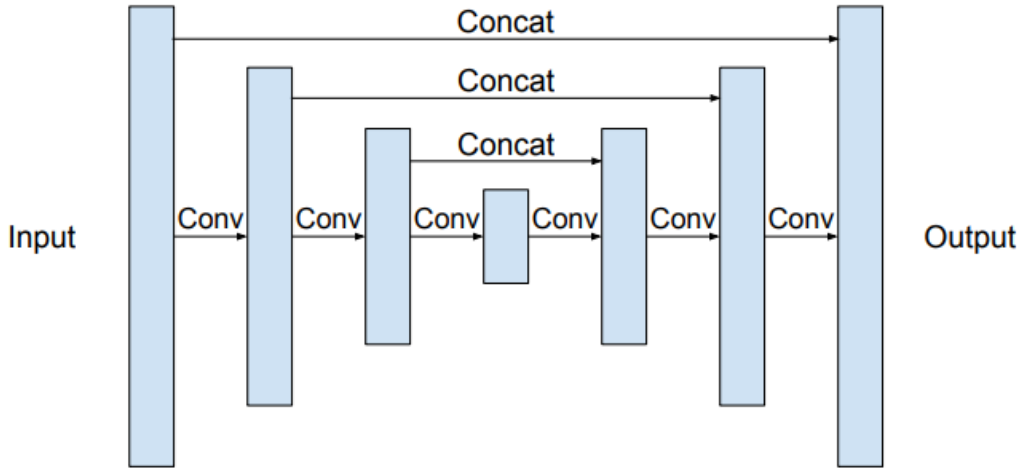

Рисунок 2.3 – Огляд архітектури U-мережі

У цій роботі кодер у конвеєрі нанесення водяних знаків працює за принципом "спектрограма до спектрограми". Створення глибокої мережі зі збереженням деталей є складним завданням, що викликає проблеми при обробці аудіо. Невеликі зміни в STFT можуть призвести до великих відмінностей у реконструйованому аудіо. Янссон та ін. [9] використовують U-мережеву архітектуру для розділення голосу співака в аудіо. Вони припускають, що здатність U-мереж зберігати низькорівневі деталі в зображеннях робить їх добре придатними для відтворення аудіо.

2.3 Оцінка водяних знаків

Модель буде навчатися окремо з різними типами шуму, а також з усіма типами шуму разом. Модель, навчена без шуму, буде використовуватися як базова. Оцінка базується на двох критеріях, один з яких вимірює спотворення несучої, а інший – частоту помилок у декодованому повідомленні. Частота помилок у декодованому повідомленні задається двійковою точністю декодованого повідомлення, яка представлена двійковим рядком. Двійкова точність A задається формулою

$$A = \frac{1}{N} \sum \left| M_{orig} - M_{pred} \right|, \quad (2.1)$$

де M_{orig} – правильне повідомлення;

M_{pred} – передбачене повідомлення;

N – довжина повідомлення в бітах.

Спотворення сигналу наведено у вигляді середньої абсолютної похибки. Вищі спотворення означають, що до носія було внесено більші зміни. Якщо зміни занадто великі, вони будуть помітні. Середня абсолютна похибка D визначається за формулою

$$\mathrm{D} = \frac{1}{N} \sum \left| I_{orig} - I_{enc} \right|, \quad (2.2)$$

де I_{orig} – оригінальний ЦВЗ;

I_{enc} – закодований ЦВЗ;

N – кількість точок даних.

3 ОГЛЯД ВИКОРИСТАНИХ МЕТОДІВ

3.1 Опис вхідних даних

Дані, що використовуються для навчання моделі, складаються з двох частин: аудіо та повідомлень.

Використані дані базуються на 2500 найбільш прослуховуваних треків комерційного музичного сервісу. З кожного треку використано приблизно 30-секундний кліп. Аудіо семплірується з частотою 44100 Гц, конвертується з MP3 у WAV і перетворюється зі стерео в моно. Доріжки перемішуються і діляться на 1500 навчальних доріжок, 500 перевірочних доріжок і 500 тестових доріжок. Тестові доріжки відкладаються в сторону і не використовуються під час розробки або навчання.

30-секундні кліпи конкатенуються, а потім розбиваються на сегменти по 88200 відліків кожен, що відповідає 2 секундам. Короткочасне перетворення Фур'є обчислюється з розміром вікна 2800 і довжиною кроку 1400. В результаті отримуємо ШПФ з 1401 частотним біном і 64 часовими кроками, розділеними на амплітуду і фазу. Модель використовує лише амплітуду в частотних діапазонах від 11 до 267, що відповідає частотам між приблизно 157 Гц і 4205 Гц. Це важливо для того, щоб метод працював при відтворенні та записі звуку в ефірі за допомогою побутового обладнання, наприклад, ноутбуків або смартфонів. Низькі частоти втрачаються в ефірі, а високі частоти будуть втрачені, якщо апаратне забезпечення має обмежену пропускну здатність. Решта амплітуд і фаза залишаються незмінними під час кодування. При відновленні звуку після кодування закодований звук об'єднується з незміненими частинами, як показано на рисунку 3.1. Декодер використовує те саме вікно амплітуди, що й кодер.

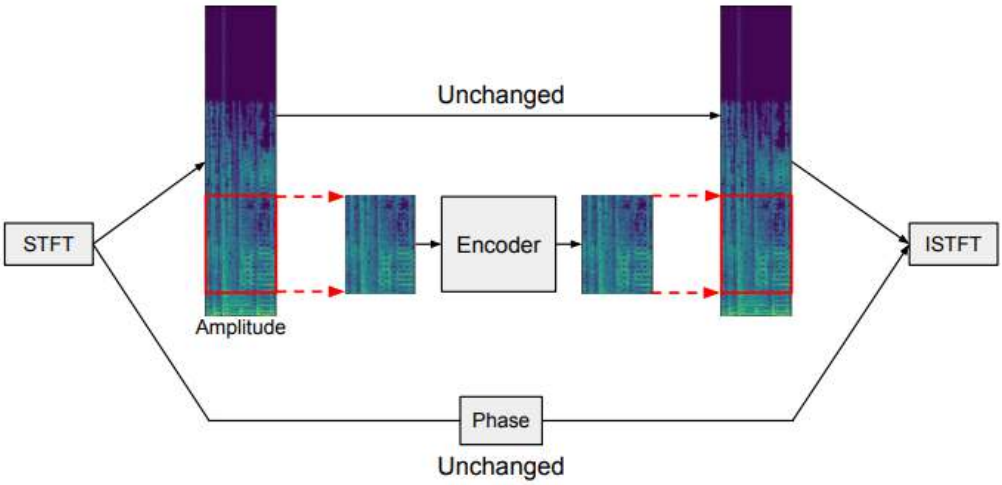

Рисунок 3.1 – Кодер обробляє лише вікно амплітуди STFT енкодером

Амплітуди нормалізуються шляхом ділення кожної вибірки xi на стандартне відхилення всієї навчальної вибірки як

$$\hat{x}_i = \frac{x_i}{std\left(x_{training}\right)} \quad , \tag{3.1}$$

де $std(x_{training})$ – середньоквадратичне відхилення вирівняних амплітуд навчальної вибірки. Ця нормалізація масштабує амплітуди, але залишає їх невід'ємними.

З навчальних, валідаційних і тестових наборів відбирається 10 % даних, які використовуються як аудіошум, як описано в розділі 3.2.3. Зразки шуму повторюються, щоб відповідати розміру відповідного набору даних.

Після відокремлення спектрограм, використаних для шуму, навчальний набір складається з 20230 спектрограм, валідаційний набір – з 6741 спектрограми, а тестовий набір – з 6744 спектрограм.

Створюються випадкові двійкові повідомлення M довжиною L відповідно до $M \in \{0, 1\}^L$. Щоб уникнути перенавчання моделі, кожна спектрограма в навчальній вибірці повторюється 5 разів з різними

повідомленнями, в результаті чого навчальна вибірка має розмір 101150. Для валідаційної та тестової множин використовується лише одне повідомлення.

3.2 Архітектура нейронної мережі

Архітектура моделі базується на трьох основних частинах: кодер, який використовується для модифікації звуку, шум, який спотворює закодований звук, і декодер, який декодує повідомлення з закодованого і спотвореного звуку.

Архітектура кодера базується на архітектурі U-net, що використовується Янссоном та ін. в [9]. Огляд архітектури кодера наведено на рисунку 3.2. Частина U-мережі, що виконує дискретизацію, складається з 5 блоків. Кожен блок містить один шар 2D згортки з кроком 2 та розміром ядра 5 × 5, пакетну нормалізацію [8] та активації Leaky ReLU [13] з негерметичністю 0.2. Після 5 блоків розмірність зменшується до 8 × 2 × 256. У цьому вузькому місці мережі повідомлення вставляється в мережу. Як і в роботі Чжу та ін. [18], повідомлення повторюється в просторовому вимірі, щоб відповідати розміру вузького місця. Повідомлення розширюється у вимірі каналу, що призводить до утворення блоку розміром 8 × 2 × 64, який візуалізовано на Рисунку 3.3. Цей блок об'єднується з вузьким місцем, після чого додається ще один шар 2D-згортки з кроком 1, розміром ядра 5 × 5 і активацією ReLU. Для відновлення даних до початкової розмірності використовується частина апсемплінгу, що складається з 5 блоків. Кожен блок складається з транспонованої 2D-згортки з кроком 2 та розміром ядра 5×5, пакетної нормалізації та активації ReLU. Перші 3 блоки мають шар відсіву з ймовірністю відсіву 50% після шару пакетної нормалізації. В кінці кожного блоку шар відповідного розміру на кроці зменшення вибірки конкатенації. Останній шар 2D-згортки з кроком 1, розміром ядра 5 × 5 і активацією ReLU використовується для зменшення кількості каналів до 1.

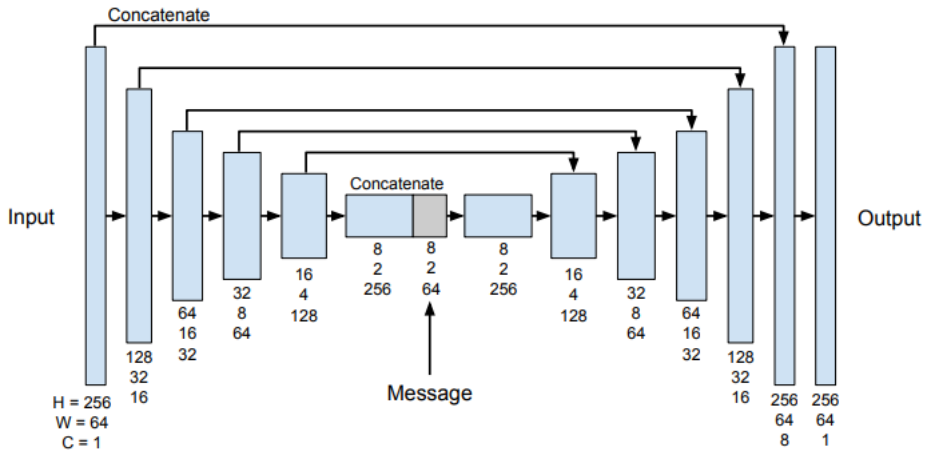

Рисунок 3.2 – Архітектура кодера з вихідними розмірами, вказаними для кожного шару

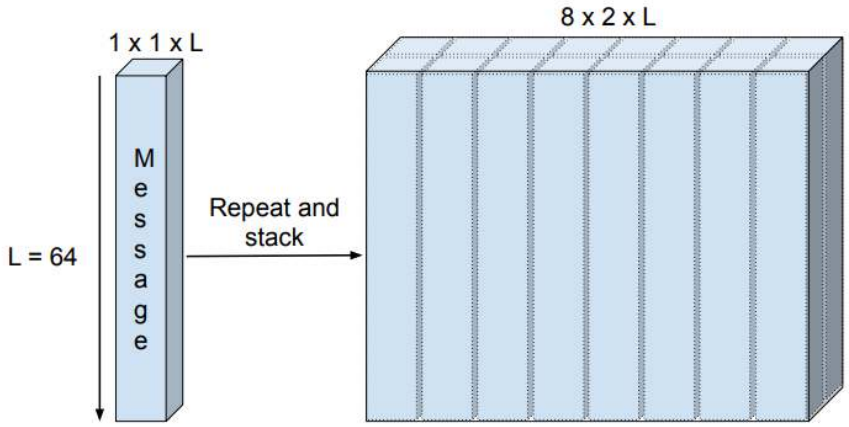

Рисунок 3.3 – Повідомлення повторюється і накопичується перед тим, як потрапити до кодера

Декодер працює як класифікатор з декількома мітками. Огляд архітектури наведено на Рисунку 3.4. Він приймає спектрограму на вхід і виводить 64 передбачення від 0 до 1, кожне з яких відповідає біту. Мережа складається з 6 блоків покрокової 2D-згортки, пакетної нормалізації та активації Leaky ReLU з коефіцієнтом витоку 0.2. Крок задано таким чином, що просторова розмірність після останньої згортки становить 64x1. Останній шар є повністю зв'язним шаром з 64 вихідними нейронами і сигмоїдними

активаціями. Виходи будуть між 0 та 1, а передбачений біт генерується шляхом округлення вихідного значення. Втрати декодера задаються двійковою перехресною ентропією між прогнозами на виході мережі та закодованим повідомленням.

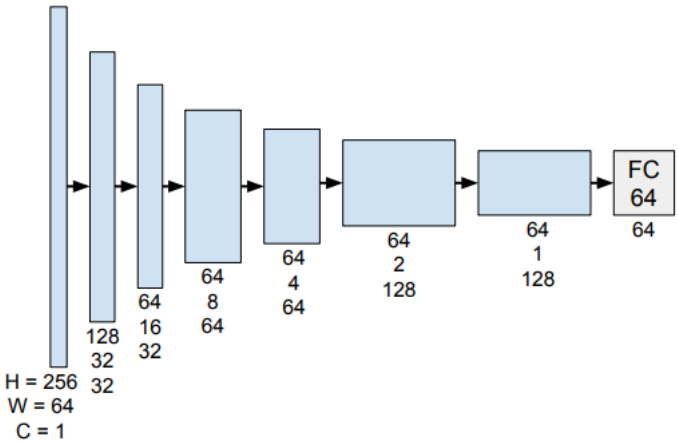

Рисунок 3.4 – Архітектура декодера з вихідними розмірами, вказаними для кожного шару

Між кодером і декодером до звуку застосовується кілька різних типів адитивного шуму. Всі типи шуму додаються безпосередньо до спектрограм звуку. Для того, щоб мати змогу включати шум при навчанні моделі, моделі шуму повинні бути диференційованими. Для цього використовуються різні типи шуму:

- -відкидання 50% пікселів спектрограми, вибраних випадковим чином, встановлюються в 0. Це призводить до надлишковості у представленні закодованого повідомлення;

- гауссівський шум До спектрограми додається гауссівський шум із середнім значенням 0 і стандартним відхиленням 0.2. Мета цього шуму – додати стійкості до загального шуму, що вноситься в ефірний канал;

- фонова музика Поверх спектрограми додається спектрограма окремого аудіокліпу. Спектрограма шуму масштабується з коефіцієнтом 0,3. Це імітує фоновий шум в ефірному каналі;

- roll & crop Спектрограма зсувається вздовж часової осі. На кордонах спектрограма загортається, так що частина спектрограми, яка зсунута вправо, додається на початку. Після згортання спектрограма обрізається на 50% вздовж часової осі. Зсув вибирається випадковим чином в межах від 0 до 64. Для того, щоб не змінювати розмір вхідного сигналу на декодері, спектрограма зсувається вправо на нуль, щоб відповідати розміру до обрізання. Мета цього шуму – зменшити необхідність узгодження сегментів між кодуванням і декодуванням. При відтворенні аудіо в ефірі записані 2-секундні сегменти не будуть синхронізовані з кодованими;

- комбінований Всі типи шуму комбінуються, впорядковані як гаусівський шум, відсікання, фонове аудіо і, нарешті, Roll & Crop. Метою використання всіх типів шуму разом є створення моделі, яка є стійкою до шуму, що вноситься під час відтворення та запису звуку в ефірі.

3.3 Навчання

Модель навчається за допомогою оптимізатора Адама [10] зі стандартними параметрами, за винятком швидкості навчання, яку зменшено до 0,0002. Модель навчається протягом 200 епох з розміром партії 128.

Функція втрат визначається як сума втрат кодера і втрат декодера, які зважуються і масштабуються для покращення збіжності. Середня абсолютна помилка у втратах кодера і двійкова перехресна ентропія у втратах декодера не мають очевидного зв'язку, що зумовлює необхідність масштабування втрат для забезпечення хорошого балансу між ними. Для моделі легко зіставити вхід кодера з його виходом, що робить його збіжність швидшою, ніж у декодера.

Щоб протидіяти цьому, на ранніх стадіях навчання втрати декодера повинні мати більшу вагу. Ваги втрат оновлюються динамічно з метою зменшення відносної різниці між втратами.

Дві цільові ваги, wencoder і wdecoder, встановлюють бажаний баланс втрат. Втрати L множаться на вагу w, щоб отримати коефіцієнт масштабування Lweighted = wL. На початку навчання ці значення коригуються, щоб надати більшу вагу втратам декодера, що дозволяє моделі йти на компроміси з реконструкцією звуку, щоб зробити можливим вбудовування представлення повідомлення. Спочатку wencoder = 1.0 і wdecoder = 2.0. Наприкінці кожної з перших 10 епох wencoder збільшується на 0.1, а wdecoder зменшується на 0.1. Після 10 епохи wencoder = 2.0 і wdecoder = 1.0, що робить більший акцент на якості реконструкції кодера

Другий метод використовується для оновлення ваг втрат в середині епохи, щоб змусити втрати бути збалансованими відповідно до цільових ваг. $\beta_{encoder}$ і $\beta_{decoder}$ – це середні втрати для останніх 10 міні-партій, масштабовані за цільовими вагами, які визначаються як

$$\beta_{encoder} = w_{encoder} * \frac{1}{10} \sum_{i=0}^{9} L_{encoder,\, b-i} \tag{3.2}$$

і

$$\beta_{decoder} = w_{decoder} * \frac{1}{10} \sum_{i=0}^{9} L_{decoder,\, b-i} \ , \tag{3.3}$$

тут $L_{encoder, n}$ та $L_{decoder, n}$ – втрати кодера та декодера після міні-партії n відповідно, b – поточна міні-партія.

Використовуючи середні втрати, вага втрат α оновлюється кожні 10 міні-партій як

$$\alpha = \frac{\beta_{encoder}}{\beta_{encoder} + \beta_{decoder}} \ . \tag{3.4}$$

4 РЕЗУЛЬТАТИ ЕКСПЕРЕМЕНТУ

4.1 Оцінка роботи методу

Для оцінки якості закодованої та реконструйованої спектрограм розраховано середню абсолютну похибку між оригінальною та закодованою спектрограмами, як описано в розділі 2.3. На рисунку 4.1 і в таблиці 4.1 представлено результати для оцінених моделей. Значення можна порівняти зі стандартним відхиленням спектрограм, яке приблизно дорівнює 1. Моделі, навчені без шуму або з одним типом шуму, мають дуже низьку середню абсолютну похибку, що відповідає невеликим модифікаціям закодованої спектрограми. Модель, навчена на всіх типах шуму, має більшу похибку.

Таблиця 4.1 – Середні абсолютні похибки для спектрограм, закодованих кожною моделлю

Тип шуму Середня абсолютна похибка	Середня абсолютна похибка
Без шуму	0.00097
Гауссівський	0.00476
Crop & Roll	0.00381
Випадання	0.00761
Музика	0.00220
Комбінований	0.

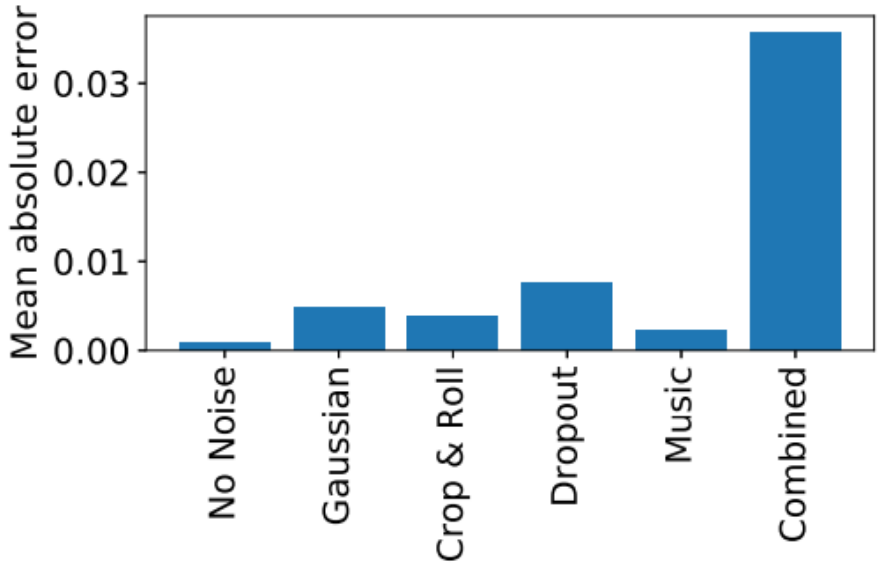

Рисунок 4.1 – Середні абсолютні похибки для спектрограм, закодованих кожною моделлю

Для оцінки робастності моделей обчислюється двійкова точність декодованого повідомлення з різними типами та кількістю шуму, що додається між кодером та декодером.

Всі моделі оцінюються без урахування шуму, що вноситься між кодером і декодером. Результати представлено на рисунку 4.2 і в таблиці 4.2. Результати показують, що всі моделі, окрім тієї, що навчена на всіх типах шуму, досягають майже 100% точності розпізнавання двійкових символів. Модель, навчена на всіх типах шуму, працює дещо гірше

Для подальшої оцінки робастності всі моделі оцінюються з використанням усіх типів шуму разом. Рівні шуму, що використовуються, такі ж самі, як і для навчання.

Результати показано на рисунку 4.3 і в таблиці 4.3, і вони показують, що всі моделі, крім тієї, що навчалася на всіх типах шуму, не досягають точності, вищої за випадковість. Модель, навчена на всіх типах шуму, досягає двійкової точності вище 99%.

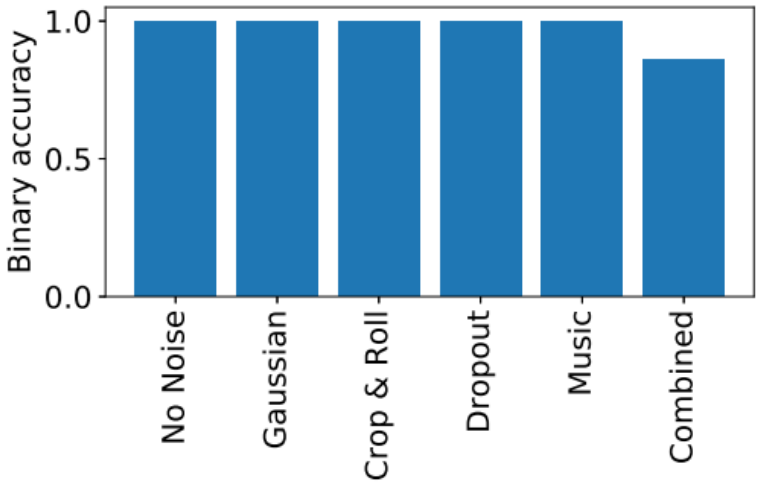

Рисунок 4.2 – Бінарна точність для всіх моделей при навчанні з типом шуму, зазначеним на вісі x, та оцінка без шуму

Таблиця 4.2 – Бінарна точність для всіх моделей при навчанні з типом шуму та оцінка без шуму

Тип шуму	Бінарна точність
Без шуму	99.991%
Гауссівський	99.990
Crop & Roll	99.965% (обрізка та роллінг)
Випадання	99.987
Фон	99.998
Комбінований	85.7

Одним з типів шуму, що використовується, є доданий гаусівський шум. Щоб оцінити вплив цього шуму на робастність, базову модель, навчену без шуму, модель, навчену лише з гауссівським шумом, і модель, навчену з усіма типами шуму разом узятими, оцінюються з гауссівським шумом. На рисунку 4.4 показано двійкову точність як функцію середньоквадратичного відхилення шуму. Моделі, навчені лише на гаусовому та комбінованому шумі, мають схожі характеристики, за винятком дуже низьких рівнів шуму, де модель з комбінованим шумом має гірші характеристики. Модель, навчена

без шуму, демонструє значне зниження двійкової точності навіть при низьких рівнях шуму.

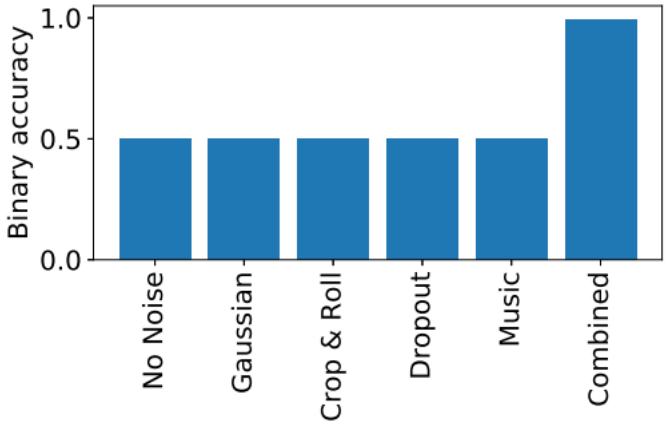

Рисунок 4.3 – Двійкова точність для всіх моделей при навчанні з типом шуму, вказаним на осі абсцис, та оцінці з усіма типами комбінованого шуму

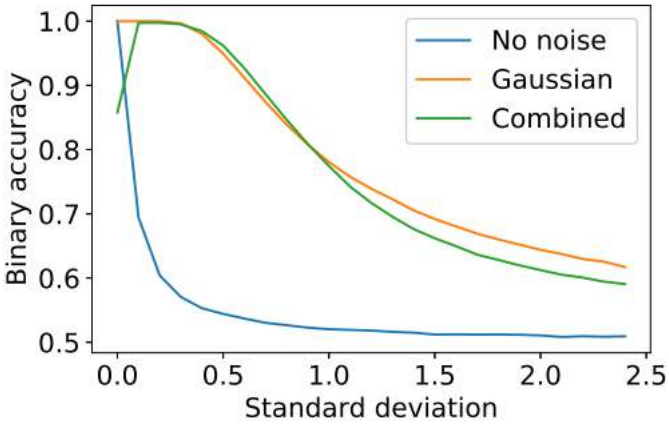

Рисунок 4.4 – Двійкова точність для моделей, оцінених за допомогою гаусівського шуму з різним стандартним відхиленням

Для оцінки стійкості до шуму, що випадає, моделі, навчені без шуму, лише з шумом, що випадає, та з усіма типами шуму, оцінюються з різною ймовірністю випадань.

Рисунок 4.5 показує двійкову точність як функцію ймовірності відсіву. Модель, навчена на випадінні шуму, має високу двійкову точність до ймовірності близько 0,7. Модель, навчена на комбінованому шумі, має гірші показники, але вищі за випадковість. Модель, навчена без шуму, стає не кращою за випадкову навіть для низьких ймовірностей.

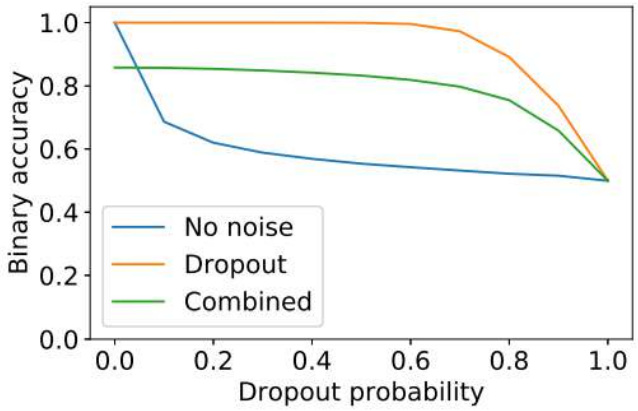

Рисунок 4.5 – Бінарна точність для моделей, оцінених з різною ймовірністю виключення

Щоб оцінити вплив шуму Roll & Crop на робастність, моделі, навчені без шуму, лише з шумом Roll & Crop, та моделі, навчені на всіх типах шуму, оцінюються з шумом Roll & Crop та без нього. Результати наведено на рисунку 4.6, з яких видно, що додавання Crop & Roll не впливає на продуктивність моделей, навчених з усіма типами шуму і з шумом Crop & Roll. Комбінована модель має нижчу двійкову точність, ніж спеціалізована модель. Модель, навчена без шуму, має двійкову точність, близьку до 100%, коли оцінюється без шуму, але працює не краще, ніж випадкова величина, коли застосовується шум

Останнім типом шуму, який було використано та оцінено, є шум фонової музики. На рисунку 4.7 показано результати для моделей, коли до закодованого аудіо додається ще одна музична доріжка. Масштаб - це коефіцієнт, на який помножено спектрограму шуму перед додаванням його до закодованого аудіо. Результати показують, що доданий музичний шум має

незначний вплив на результат. Всі моделі показують хорошу продуктивність, навіть коли масштаб перевищує 0.3, що використовується для навчання. Модель, навчена без шуму, має дещо нижчу продуктивність при оцінці без шуму.

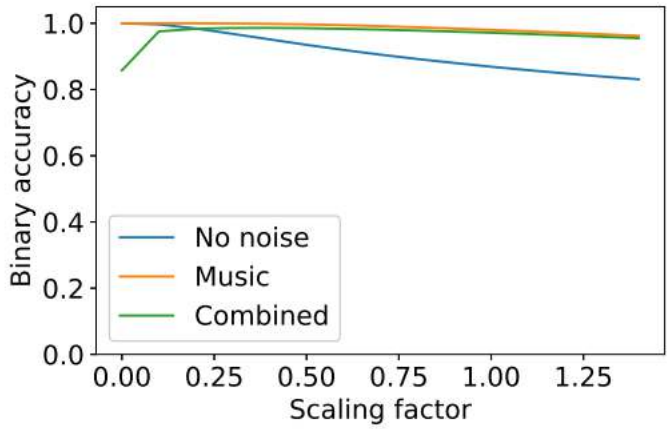

Рисунок 4.7 – Бінарна точність для моделей при оцінці музичного шуму з різним масштабуванням

Щоб оцінити, чи можливо витягти закодоване повідомлення після відтворення та запису аудіо в ефірі, ми відтворюємо та записуємо кілька треків в ефірі у фізичному світі.

З тестового набору випадковим чином вибирається 10 треків. Кожен з треків розбивається на 14 блоків по 88200 відліків кожен, кодується за допомогою моделі, навченої на всіх типах шуму разом узятих, а потім знову збирається разом. Всі 14 блоків кодуються одним і тим же повідомленням. Потім треки відтворюються з динаміка ноутбука і записуються за допомогою портативного смартфона. Записаний звук попередньо обробляється так само, як і тестові дані, а потім подається на декодер. Оскільки дані обрізаються, для кожної доріжки можна згенерувати приблизно 28 прогнозів. Середня двійкова точність цих прогнозів наведена в таблиці 4.4. Двійкова точність становить від 40% до 60% для всіх моделей, а середній показник близький до 50%.

На рисунку 4.8 показано гістограму розподілу кількості правильних бітів для всіх 279 прогнозів. Розподіл зосереджено навколо 32 бітів.

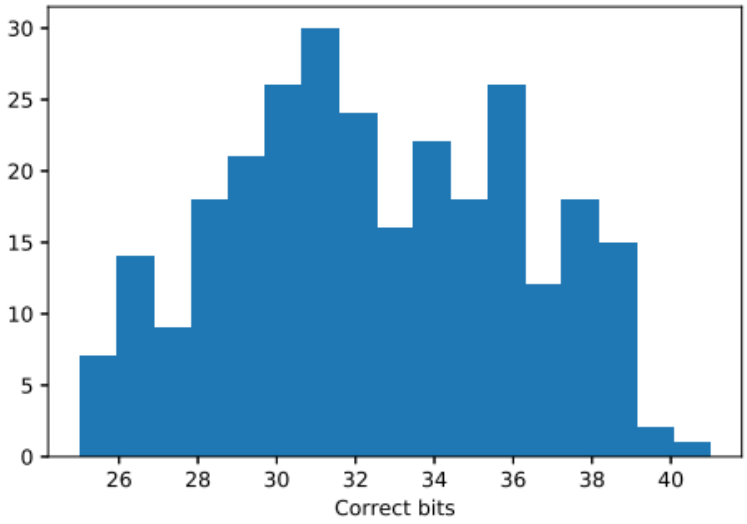

Рисунок 4.8 – Розподіл кількості правильних бітів для всіх прогнозів

Таблиця 4.4 – Середня бінарна точність для кожної з 10 доріжок, використаних для оцінювання.

Треки	Правильні біти (в середньому)	Відсоток
Track 1	30.92	48.32%
Track 2	31.0	48.43%
Track 3	29.46	46.03%
Track 4	28.21	44.08%
Track 5	36.35	56.80%
Track 6	37.96	59.31%
Track 7	35.18	54.97%
Track 8	35.10	54.85%
Track 9	32.14	50.22%
Track 10	28.07	43.86%
All	32.43	50.67%

Для візуалізації того, як повідомлення представлено в закодованих спектрограмах, для різних моделей, повідомлень і спектрограм було побудовано графіки залишків між закодованими і вихідними спектрограмами.

Для візуалізації того, як повідомлення представлені в закодованих спектрограмах для різних моделей, закодовані спектрограми X одного тестового зразка показані на рисунку 4.9. Для більшої наочності спектрограми побудовано як log(1 + X)

Абсолютне значення залишку між закодованою спектрограмою та вихідною спектрограмою X показано на рисунках 4.10 та 4.11, де залишки було масштабовано до log(1 + 10|X - X|) та log(1 + 50|X - X|) відповідно. Залишки показують чіткі відмінності в тому, як представлені повідомлення для різних моделей

Вихід кодера залежить як від вхідної спектрограми, так і від повідомлення, що кодується. Щоб візуалізувати, як ці два фактори впливають на вихід кодера, дві різні спектрограми кодуються одним і тим же повідомленням. Щоб побачити, як зміна повідомлення впливає на вихід, перша спектрограма кодується двійковим доповненням першого повідомлення.

Для цього тесту використовується модель, навчена на всіх типах шуму разом узятих. Залишки, побудовані як log(1 + 10|X - X|), показано на рисунку 4.12. Дві спектрограми, закодовані тим самим повідомленням, мають схожий вигляд залишків, тоді як третя спектрограма, закодована двійковим доповненням вихідного повідомлення, виглядає інакше

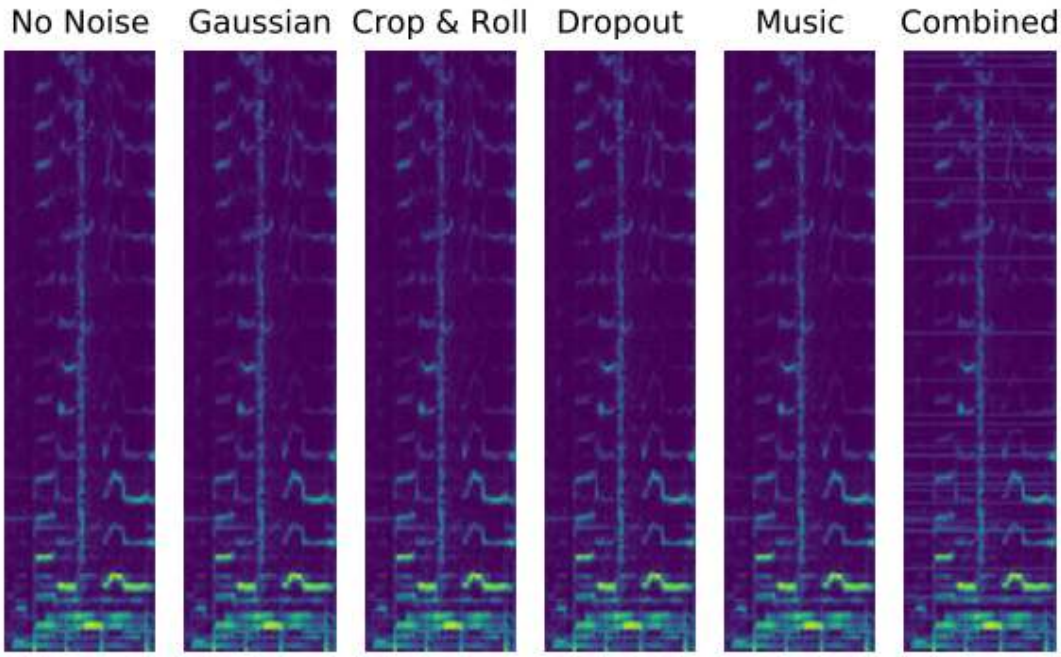

Рисунок 4.9 – Кодовані спектрограми для всіх моделей

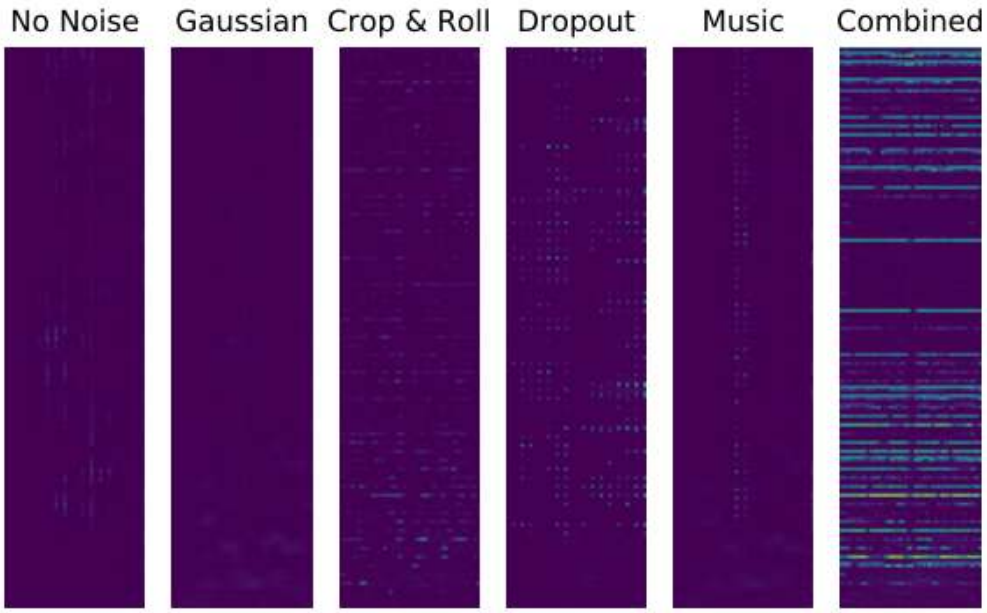

Рисунок 4.10 – Залишки для закодованих спектрограм, масштабовані з коефіцієнтом 10

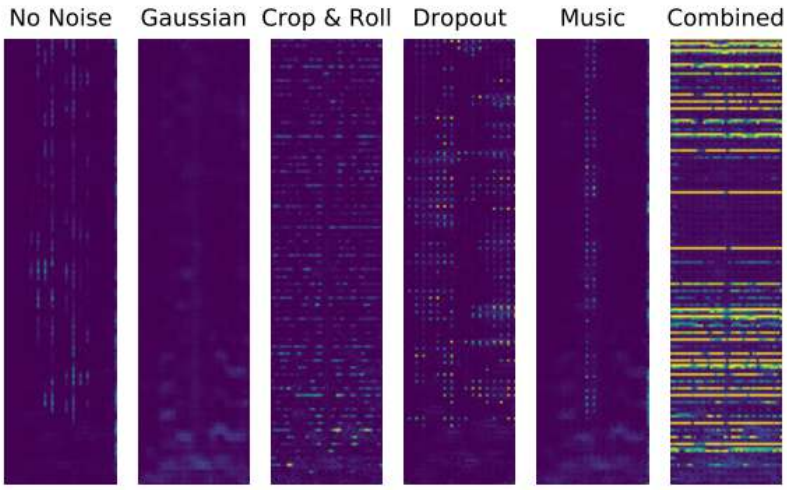

Рисунок 4.11 – Залишки для закодованих спектрограм, масштабовані з коефіцієнтом 50

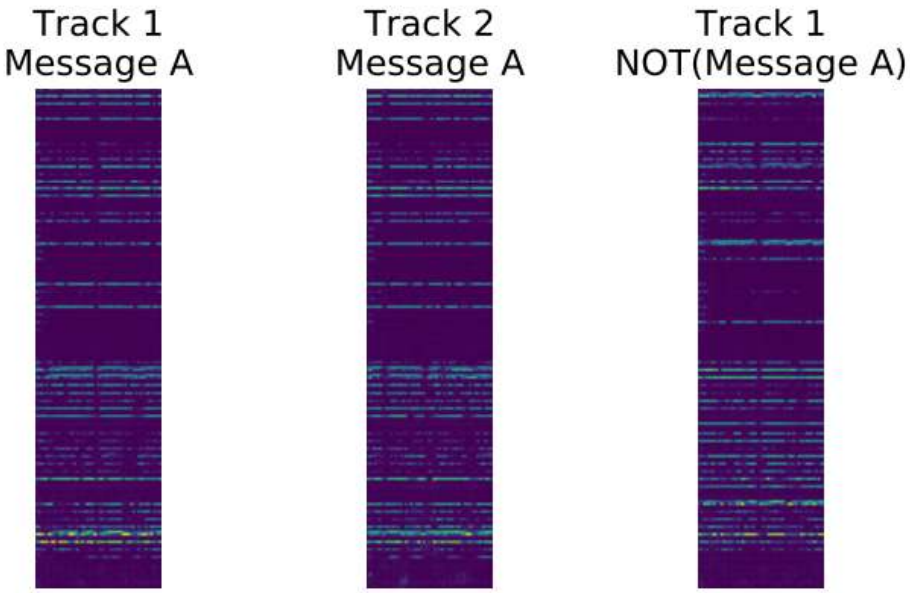

Рисунок 4.12 – Залишки двох спектрограм, закодованих одним і тим же повідомленням, та залишок лівої спектрограми, закодованої двійковим доповненням повідомлення

4.2 Обговорення результатів

Результати показують, що система здатна кодувати і витягувати повідомлення з усіма типами шуму. При навчанні лише на одному типі шуму, або без шуму, декодування може бути виконано з дуже високою стійкістю і лише з невеликими модифікаціями несучої.

З іншого боку, коли всі типи шуму комбінуються, кодування все ще є успішним, але зниження якості звуку стає значним. При прослуховуванні кодованого аудіо можна почути статичний фоновий шум, особливо якщо в аудіо носія є тихі ділянки.

Оцінюючи модель з тією ж кількістю і типом шуму, на якому вона навчалася, можна сказати, що вона має дуже високу робастність. Цікаво, однак, що модель, навчена на всіх типах шуму разом узятих, працює гірше, коли її оцінюють на меншій кількості шуму, наприклад, коли використовується лише один тип шуму або без шуму. Це вказує на те, що модель могла створити залежність від шуму для вилучення повідомлення, подібно до надмірної підгонки. Це може знизити продуктивність при відтворенні в ефірі, якщо змодельований шум не відповідає реальному шуму.

При оцінці моделей з використанням музичного шуму видно, що всі оцінені моделі працюють досить добре навіть з масштабним коефіцієнтом, більшим за той, що використовувався для навчання, навіть модель, навчена без шуму. Це вказує на те, що даний тип шуму може не підходити для даної задачі. Випадково вибрані треки, що використовуються при навчанні, можуть бути дуже слабкими і загалом непередбачуваними.

Однією з цілей цієї роботи є підвищення стійкості водяних знаків до шуму. Було використано та оцінено різні типи шуму, але, як видно з моделі, навченої на всіх типах шуму, не очевидно, що моделі працюють добре, коли шум відрізняється від того, що використовувався при навчанні. Результати для моделі, навченої з шумом відсіву, показують, що ймовірність відсіву може бути збільшена до рівня, вищого за той, що використовувався для

навчання, з відносно невеликим зниженням двійкової точності. Те ж саме можна побачити для моделі, навченої на гаусівському шумі на рисунку 4.4. Ці типи шуму створюють моделі, менш чутливі до кількості шуму в каналі.

Дивлячись на залишки закодованих повідомлень на рисунку 4.12, стає зрозуміло, що кодування в основному базується на повідомленні, а не на спектрограмі. Різні спектрограми дають лише невеликі зміни, але зміна повідомлень має велике значення. Той факт, що представлення повідомлення не робить більших адаптацій при зміні спектрограми, призводить до чутних помилок у реконструйованому аудіо. Можливо, що більш адаптивне і динамічне представлення краще приховувало б повідомлення.

Тест при відтворенні аудіо в ефірі показує середню бінарну точність, близьку до 50%, тобто не кращу за випадковий вибір. Якщо поглянути на окремі треки, то деякі з них досягають дещо вищої двійкової точності, близької до 60%. Це свідчить про певний потенціал, але для досягнення надійних результатів потрібно докласти більше зусиль. Можливо, деяких поліпшень можна досягти за рахунок кращої попередньої обробки записаних даних, наприклад, за рахунок зашумлення і нормалізації гучності.

Очевидно, що кодеру вдається вбудувати повідомлення в аудіо, зберігаючи при цьому деталі в аудіо. Дивлячись на залишки спектрограм закодованого аудіо, представлення повідомлення не здається дуже складним. Воно додає або видаляє енергію в деяких діапазонах частот. Виникає питання, чи є сенс використовувати таку складну модель з понад 5 мільйонами ваг для створення системи з таким простим представленням. Модель не була оптимізована щодо гіперпараметрів через брак часу. Цілком можливо, що краще оптимізована модель могла б працювати краще або досягти подібних результатів з меншою кількістю ваг.

Дивлячись на залишки на рисунку 4.11, видно, що шум Crop & Roll вносить лінії в залишки. При поєднанні цього шуму з іншими типами результат показує помітні лінії вздовж часової осі, створюючи статичний шум на тлі вихідного аудіо. Якби проблему часової синхронізації можна було

вирішити іншим способом, ніж додаванням шуму Crop & Roll, можна було б досягти кращих результатів при менших витратах на зменшення якості.

Іншим фактором, який може вплинути на просте представлення повідомлень, є вибір метрик якості для кодера. Середня абсолютна похибка однаково карає малі та великі помилки, роблячи великі помилки можливими. У той же час розумно вважати, що декілька великих помилок є більш стійкими до шуму, ніж декілька малих помилок. Середня абсолютна похибка не обов'язково відповідає сприйнятій якості звуку, і можливим напрямком покращення може бути використання кращої функції втрат, наприклад, психоакустичної моделі, використаної в роботі Цинь та ін. [14], яка була опублікована ближче до завершення цієї кваліфікаційної роботи.

Альтернативним підходом до створення більш глибокого представлення, яке є менш очевидним, може бути додавання супротивника до моделі, як це зроблено в роботі Чжу та ін. [18]. Чіткі лінії в залишках, отриманих нашою моделлю, ймовірно, будуть виявлені і покарані супротивником. Ця гіпотеза також підтверджується роботою [18], де викривлення набагато помітніші, якщо не включати супротивника в модель.

Недоліком такого підходу є збільшення проблем зі збіжністю. Противник додає третю втрату, яка повинна бути збалансована з двома втратами в поточній моделі. Можна було б дослідити більш складні методи балансування втрат, такі як метод градієнтних нормалізацій, запропонований Ченом та ін. [4].

Іншою сферою вдосконалення є моделювання шуму. Результати показують, що модель має тенденцію до надмірної адаптації до шуму, який використовується в навчанні. Простим рішенням може бути зміна кількості шуму, що використовується під час навчання, але є також можливості для значних покращень за рахунок використання більш реалістичного шуму. Існуючі методи добре підходять для введення надмірності в представлення повідомлень, але більш реалістичний шум, такий як акустична симуляція

кімнати, яку використовували Цинь та ін. [14], або реакції мікрофона і динаміка, може працювати краще при відтворенні в ефірі.

Оцінка продуктивності при відтворенні в ефірі виконується на невеликому наборі треків лише при одному налаштуванні обладнання та оточення. Це обмежує можливість робити якісь сильні припущення щодо продуктивності в загальному випадку. Для отримання порівнянних результатів необхідно провести більш ретельне оцінювання.

ВИСНОВКИ

У цій кваліфікаційній роботі запропоновано метод нанесення водяних знаків на аудіо з використанням глибокого навчання. Метод складається з двох глибоких нейронних мереж, кодера і декодера, які були навчені наскрізним навчанням з використанням музики та імітованого шуму. Базова модель тренується без імітованого шуму і порівнюється з моделями, навченими з використанням різних типів шуму, таких як відсікання, гаусівський шум, фонова музика, обрізка і рол, а також комбінація цих типів шуму. Метод добре працює, коли використовується лише один тип шуму, коли повідомлення можна виділити з високою надійністю і невеликим зниженням якості. Моделі, навчені з імітованим шумом, перевершують базову модель, навчену без шуму, щодо стійкості до імітованого шуму. При використанні декількох типів шуму модель також здатна витягти закодоване повідомлення, але зниження якості звуку є більшим, і модель не дуже добре узагальнює кількість шуму. Незважаючи на те, що модель досить стійка до імітованого шуму, в реальних умовах неможливо виділити закодоване повідомлення після відтворення і запису закодованого аудіо в ефірі. Результати також свідчать про те, що модель не вбудовує повідомлення в аудіо дуже складним способом. Кодування не робить великих адаптацій до різних спектрограм, і тому помилки, що виникають в результаті, призводять до статичного шуму в закодованому аудіо.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Shumeet Baluja. Hiding images in plain sight: Deep steganography. In Neural Information Processing Systems, 2017.

2. Walter Bender, Daniel Gruhl, Norishige Morimoto, and Anthony Lu. Techniques for data hiding. IBM systems journal, 35(3.4):313–336, 1996.

3. Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. arXiv preprint arXiv:1801.01944, 2018.

4. Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. arXiv preprint arXiv:1711.02257, 2017.

5. Ingemar J Cox, Joe Kilian, F Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. IEEE transactions on image processing, 6(12):1673–1687, 1997.

6. Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations, 2015. URL http://arxiv.org/abs/1412.6572.

7. Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning, volume 1. MIT press Cambridge, 2016.

8. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. pages 448–456, 2015.

9. Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep u-net convolutional networks. 2017

10. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

11. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

12. Yiqing Lin and Waleed H Abdulla. Principles of psychoacoustics. In

Audio Watermark, pages 15–49. Springer, 2015.

13. Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In Proc. icml, volume 30, page 3, 2013.

14. Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. arXiv preprint arXiv:1903.10346, 2019.

15. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

16. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

17. Ron G Van Schyndel, Andrew Z Tirkel, and Charles F Osborne. A digital watermark. In Proceedings of 1st International Conference on Image Processing, volume 2, pages 86–90. IEEE, 1994.

18. Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 657–672, 2018.

19. Гунько, Максим, et al. " Метод вбудови цифрових водяних знаків в аудіо-файли за допомогою машинного навчання." InterConf (2023): 210-217.