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)

Моделі та методи підвищення аудіоякості музичних творів із використанням нейронних мереж

(тема)

Виконав:
здобувач 2 року навчання,
групи СПм-23-5
Максим ЖУК
(власне ім'я, прізвище)

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

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

Освітня програма
Системне програмування
(повна назва освітньої програми)

Керівник: проф. Тетяна ФЕСЕНКО
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ЕОМ _____ Андрій КОВАЛЕНКО
(підпис) (власне ім'я, прізвище)

2025 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-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

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

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Жуку Максиму Володимировичу
(прізвище, ім'я, по батькові)

1. Тема роботи Моделі та методи підвищення аудіоякості музичних творів із використанням нейронних мереж

затверджена наказом по університету від "21" квітня 2025 р. № 296 СТ

2. Термін подання здобувачем роботи до екзаменаційної комісії 16 червня 2025 р.

3. Вхідні дані до роботи 1) Операційна система Windows; 2) Python; 3) IDE PyCharm; 4) Google Colab; 5) SQL-бази даних; 6) PyTorch; 7) Torchaudio; 8) бібліотека librosa; 9) Pandas; 10) NumPy; 11) Matplotlib; 12) публічні датасети MusDB18; 13) Kaggle; 14) Free Music Archive; 15) інтерфейси для прослуховування та візуалізації аудіо; 16) GitLab; 17) хмарні сервіси для зберігання даних; 18) існуючі open-source рішення

4. Перелік питань, що потрібно опрацювати у роботі
1) аналіз сучасних підходів до оцінки та підвищення аудіоякості;
2) огляду існуючих рішень та архітектур;
3) вибор цільових метрик;
4) підготовки аудіоданих і середовища розробки;
5) побудови та навчання багатоступеневої нейромережевої моделі;
6) експериментального дослідження
7) ефективності та порівняння результатів із сучасними підходами

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

Слайд-презентація – 12 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	Строк / терміни виконання етапів роботи	Примітка
1	Постановка задачі, вибір метрик якості	22.04.25-29.04.25	
2	Підбір та підготовка аудіоданих	30.04.25-05.05.25	
3	Розробка та тестування нейромережевої	06.05.25-18.05.25	
	архітектури		
4	Побудова пайплайну обробки	19.05.25-24.05.25	
5	Проведення навчання моделі, експеримен-	25.05.25-02.06.25	
	тальна перевірка, аналіз		
6	Оформлення матеріалів кваліфікаційної роботи	03.06.25-05.06.25	
7	Подання кваліфікаційної роботи керівникові	06.06.25-09.06.25	
	та її попередній захист		
8	Подання кваліфікаційної роботи на рецензу-	10.06.25-12.06.25	
	вання		

Дата видачі завдання " 21 " квітня 2025 р.

Здобувач _____ (підпис)

Керівник роботи _____ (підпис) проф. Тетяна ФЕСЕНКО (посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 92 с., 32 рис., 6 табл., 1 дод., 22 джерел.

АУДІОСИГНАЛ, ШУМОЗАГЛУШЕННЯ, СПЕКТРОГРАМА, НЕЙРОННА МЕРЕЖА, ЗГОРТКА, STFT, WAV, ЯКІСТЬ, SNR, PYTHON.

Основною метою кваліфікаційної роботи є дослідження моделей та методів підвищення аудіоякості музичних творів із використанням нейронних мереж. Спочатку проведено ґрунтовний огляд сучасних підходів до оцінки та поліпшення аудіосигналів, серед яких виділено класичні спектральні методи та передові нейромережеві рішення. Аналіз показав, що традиційні алгоритми часто нездатні адаптуватися до широкого різноманіття жанрових особливостей музики та різного рівня шумового фону, тоді як нейронні мережі демонструють високу здатність до навчання комплексних патернів та більш гуманізованого відновлення звуку.

У ході виконання кваліфікаційної роботи було виконано огляд сучасних методів оцінки та покращення аудіоякості, на основі якого сформовано корпус аудіоданих із музичних, мовних та шумових фрагментів. Запропоновано багатоетапну архітектуру, що поєднує грубе шумоподавлення на основі ERB-масок, спектральну реконструкцію трансформером та GAN-постобробку з контекстним маскуванням і фазовим коректором. Впроваджено адаптивний механізм маршрутизації обробки залежно від проміжної оцінки якості сигналу. Прототип реалізовано на Python із використанням бібліотек PyTorch, Librosa та Torchaudio, і експериментально підтверджено покращення показників PESQ, SI-SDR та ViSQOL порівняно з класичними моделями.

ABSTRACT

Master’s thesis: 92 pages, 32 figures, 6 tables, 1 appendices, 22 sources.

AUDIO SIGNAL, NOISE SUPPRESSION, SPECTROGRAM, NEURAL NETWORK, CONVOLUTION, STFT, WAV, QUALITY, SNR, PYTHON.

The major goal of this thesis is to investigate and develop effective models and methods for enhancing the audio quality of musical works using neural networks. Achieve this, a comprehensive review of contemporary audio-processing techniques was carried out, encompassing classical spectral methods such as STFT-based filtering and convolutional denoising, alongside cutting-edge deep-learning architectures.

In order to evaluate and compare these approaches, a heterogeneous corpus of WAV recordings was assembled, incorporating diverse musical genres, speech samples, and synthesized noise at varying SNR. A multi-stage neural architecture was then proposed, beginning with a coarse noise-suppression module based on ERB masking, followed by a transformer-based block for detailed spectrogram reconstruction, and concluding with a GAN-driven post-processing stage featuring convolutional layers and phase correction to eliminate residual artifacts. An adaptive routing mechanism was implemented to dynamically select processing paths based on intermediate quality estimates, thereby optimizing performance across different noise profiles and spectral complexities. The prototype, realized in Python using PyTorch, Librosa, and Torchaudio, demonstrated substantial gains in objective metrics–PESQ, SI-SDR, and ViSQOL when benchmarked against classical two-stage models.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

dB – децибел (англ., decibel)

Гц – герц (англ., hertz, скорочено Hz)

AES – Товариство інженерів аудіо (англ., Audio Engineering Society)

CNN – згорткова нейронна мережа (англ., Convolutional Neural Network)

CUDA – уніфікована архітектура обчислень для пристроїв NVIDIA (англ., Compute Unified Device Architecture)

DCCRN – глибока комплексна згорткова рекурентна мережа (англ., Deep Complex Convolution Recurrent Network)

DNSMOS – об'єктивна оцінка якості мовлення на основі нейронної мережі (англ., Deep Noise Suppression Mean Opinion Score)

DVC – диференційоване керування гучністю (англ., Dynamic Volume Control)

ERB – еквівалентна прямокутна смуга пропускання (англ., Equivalent Rectangular Bandwidth)

FLAC – безвтратний аудіоформат стиснення (англ., Free Lossless Audio Codec)

GAN – змагальна генеративна мережа (англ., Generative Adversarial Network)

GPU – графічний процесор (англ., Graphics Processing Unit)

GRU – рекурентна нейронна мережа з елементом гейтування (англ., Gated Recurrent Unit)

ICASSP – Міжнародна конференція з акустики, мовлення та обробки сигналів (англ., International Conference on Acoustics, Speech, and Signal Processing)

ITU – Міжнародний союз електрозв'язку (англ., International Telecommunication Union)

LSTM – довготривала короткочасна пам’ять (англ., Long Short-Term Memory)

MOS – середній бал якості мовлення (англ., Mean Opinion Score)

PESQ – об’єктивна оцінка сприйманої якості мовлення (англ., Perceptual Evaluation of Speech Quality)

POLQA – перцептивна оцінка якості мовлення (англ., Perceptual Objective Listening Quality Analysis)

RMS – середньоквадратичне значення (англ., Root Mean Square)

RNN – рекурентна нейронна мережа (англ., Recurrent Neural Network)

SDR – співвідношення сигнал/спотворення (англ., Signal-to-Distortion Ratio)

SI-SDR – масштабно-нормалізоване співвідношення сигнал /спотворення (англ., Scale-Invariant Signal-to-Distortion Ratio)

SNR – співвідношення сигнал/шум (англ., Signal-to-Noise Ratio)

STFT – короткочасне перетворення Фур'є (англ., Short-Time Fourier Transform)

U-Net – згорткова нейронна мережа з U-подібною архітектурою (англ., U-Net)

VISQOL – модель візуальної якості для об'єктивного оцінювання слухового сприйняття (англ., Virtual Speech Quality Objective Listener)

WAV PCM – формат аудіофайлів із безстисненим кодуванням за методом імпульсно-кодової модуляції (англ., Waveform Audio File Format – Pulse-Code Modulation)

ВСТУП

Проблема покращення якості аудіосигналів у складних акустичних умовах є актуальною у сфері цифрової обробки звуку. Особливо стосується музичних творів, які характеризуються багатошаровою спектральною структурою, динамічною варіативністю та високими вимогами до перцептивної цілісності. Більшість існуючих нейромережевих рішень, зокрема ті, що орієнтовані на обробку мовлення, демонструють обмежену ефективність при роботі з музичними даними через відсутність архітектурної адаптації до немовних сигналів. Створюючи потребу у створенні спеціалізованих моделей, здатних до гнучкої обробки складного аудіоконтенту із збереженням його гармонійної, тембрової та ритмічної структури.

Об’єктом кваліфікаційної роботи є процес цифрової обробки музичних аудіосигналів, а предметом – нейромережеві моделі та методи підвищення їх якості. Метою роботи є розробка адаптивної багатоступеневої архітектури для покращення аудіоякості музичних творів з урахуванням типу вхідного сигналу, спектральної складності та шумового профілю.

У межах кваліфікаційної роботи були поставлені наступні задачі:

- проаналізувати існуючі архітектури шумозаглушення та виявити їх обмеження при застосуванні до музичних сигналів;
- розробити методи попереднього аналізу аудіоконтенту та формування корпусу навчальних і тестових даних;
- обґрунтувати доцільність використання багатоступеневої архітектури для задач підвищення аудіоякості;
- розробити методи етапної обробки аудіосигналу, спектральне очищення, реконструкцію компонентів, зменшення артефактів;
- визначити метрики для оцінки якості сигналу та провести аналіз ефективності запропонованого підходу порівняно з базовими рішеннями.

1 ТЕОРЕТИЧНІ ЗАСАДИ ОЦІНКИ ЯКОСТІ МУЗИЧНИХ ТВОРІВ

1.1 Аналіз підходів до оцінки якості музичних творів

Оцінка якості музичних творів є складною задачею, що поєднує багато методів: цифрової обробки сигналів, психоакустики, машинного навчання та когнітивних наук. Як відзначають Valentin Emiya, Emmanuel Vincent, Niklas Harlander, Volker Hohmann у своїй роботі "Subjective and Objective Quality Assessment of Audio Source Separation" [5], якість аудіосигналу визначається, як результат взаємодії об'єктивних, вимірюваних параметрів з одного боку, і суб'єктивного сприйняття слухачів з іншого. Відповідно до їх визначенень, дослідження традиційно виділяють два основні підходи до оцінки якості музики: суб'єктивний, що базується на індивідуальному сприйнятті та оцінці слухачів, і об'єктивний, що оперує кількісними характеристиками аудіосигналу.

Розглянемо спочатку суб'єктивні методи оцінки якості музичних творів, спрямованих на особливості людського сприйняття звуку. Базуються в основному на опитуваннях слухачів, експертних оцінках професійних музикантів, критиків чи продюсерів, а також на психоакустичних дослідженнях, що вивчають реакції мозку та нервової системи на різні аудіостимули. При цьому оцінка музичних творів проводиться за такими ключовими критеріями, як гармонійність, тембральна насиченість, динаміка, ритмічна структура, емоційний вплив та відповідність жанровим очікуванням. У дослідженні А.О. Войтовича у роботі «Суб'єктивна оцінка звучання оркестрів (на прикладі концертної зали Львівської філармонії ім. С. Людкевича)» [4] автор продемонстрував, що за ретельної організації прослуховувань із залученням як професійних музикантів, так і досвідчених слухачів можна досягти високої узгодженості оцінок звучання. У результаті дослідження вдалося виділити ключові параметри, що впливають на

сприйняття оркестрового звучання, а саме: баланс інструментів, гармонійність, просторова повнота, динаміка, ритмічна структура, звуковий баланс, тембральна насиченість, шумові перешкоди, частота звучання, динамічний діапазон, емоційний вплив та відповідність жанровим очікуванням. Застосування комплексного підходу, дозволило досягти деталізованого сприйняття окремих характеристик звучання, сформувати стандартизовані критерії для суб'єктивної оцінки якості концертному залі.

Суб'єктивні методи оцінки якості звуку, незважаючи на їх здатність враховувати емоційні та художні аспекти, мають низку обмежень, оскільки сприйняття музики індивідуальне та залежить від особистого досвіду, культурного розвитку та емоційного стану слухачів, що призводить до варіативності результатів та знижує їх об'єктивність. У зв'язку з цим виникає потреба у більш точних та універсальних підходах, здатних враховувати індивідуальні відмінності. На 131-му конгресі AES в Нью-Йорку професор Чарльз Лімб з Медичної школи Університету Джона Хопкінса у своїй доповіді на тему «Звук, слух і музика: сприйняття і нейробіологія» підкреслив, що сприйняття музики є складним нейробіологічним процесом, який значно різниться між людьми, що ускладнює стандартизацію суб'єктивних оцінок якості звуку. За підсумками його виступу Чарльз продемонстрував, наскільки важливо враховувати особливості індивідуума при сприйнятті та оцінці якості звуку і як впливає на розробку об'єктивних методів аналізу.

Якщо розглядати об'єктивні методи, то навпаки, вони найбільше спираються на кількісні параметри аудіосигналу, які часто аналізуються алгоритмічними моделями, що особливо актуально у контексті сучасних досліджень у галузі аудіоякісті у музичних творів з використанням нейронних мереж [6]. Методи використовують цифрову обробку сигналів для отримання та аналізу параметрів, як спектральна щільність потужності, ступінь спотворень, баланс частот, рівень шумів, стабільність динаміки та спектральні аномалії. Вони формують основу цифрових рішень для

моніторингу, оптимізації та реконструкції аудіосигналів у мовних та музичних даних.

Наочним підтвердженням ефективності описаних методів є дослідження Sajjad Amini, Shahrokh Ghaemmaghami, яке представлено на ICASSP у роботі "Towards Perceptual Metrics for Enhanced Music Quality Evaluation" [5]. У ньому автори наголошують на обмеженості класичних методів на кшталт PESQ і POLQA в задачах, пов’язаних із музикою, оскільки вони спираються на моделі мовного сприйняття. Пропонують підхід до побудови метрик, що враховують перцептивно значущі спектральні відмінності, характерні саме для музичних сигналів, і тестують їх ефективність на даних, оброблених нейромережевими аудіомоделями. Демонструючи загальну тенденцію переходу від універсальних метрик до контент-специфічних, адаптивних систем, що здатні диференційовано оцінювати якість залежно від природи сигналу. Усе частіше такі метрики вбудовуються безпосередньо в архітектуру моделей глибокого навчання як функції втрат або валідаційні критерії. Важливо підкреслити, що для реалізації побудови систем необхідна чітко визначена метрикована база, охоплюючи параметри, включаючи рівень шуму, ступінь спотворення, динамічний та частотний діапазон, енергетичний баланс та гармонійну точність. Їх не можна розглядати ізольовано: лише в сукупності вони дозволяють створити об'єктивну основу для оцінки аудіо, незалежно від походження чи жанрової природи.

1.2 Системи та критерії оцінки якості аудіозаписів (MOS, PESQ/ POLQA, SI-SDR/ SDR, VISQOL)

Якість аудіосигналів оцінюють за двома групами критеріїв: суб’єктивними та об’єктивними. До першої належить шкала MOS; до другої – алгоритми PESQ, POLQA, SI-SDR і VISQOL. Точність та діапазон застосування кожного інструмента визначаються природою сигналу й

вимогами завдання.

MOS – це стандартний метод для суб'єктивної оцінки якості аудіосигналів, який широко використовується в галузях телекомунікації та аудіотехнології. Грунтується на оцінці сприйняття якості аудіо людьми, де учасники тесту оцінюють звучання аудіофрагменту за шкалою від 1 до 5, де 5 позначає найвищу якість, а 1 – найгіршу. Для розрахунку виставлених оцінок застосовується формула:

$$MOS = \frac{1}{N}\sum_{i=1}^{N} R_i, \tag{1.1}$$

де R_i – Оцінка, виставлена і-им учасником;

N – загальна кількість учасників тесту.

Однак важливим обмеженням цього стандарту є необхідність залучення великої кількості учасників, роблячи процес дорогим та тимчасовим. Крім того, результати можуть залежати від особистих переваг учасників та умов прослуховування обладнання та акустичних приміщень. Для підвищення автоматизації оцінки та зниження витрат було розроблено об'єктивні методи: PESQ та POLQA, які дозволяють проводити оцінку якості без участі людини.

PESQ був розроблен ITU для об'єктивної оцінки якості промови. Він є методом повного порівняння, тобто для роботи потрібно вихідний еталонний сигнал для порівняння зі спотвореною версією аудіофайлу. PESQ аналізує різницю між цими сигналами, використовуючи модель сприйняття, яка намагається імітувати сприйняття людиною різних видів спотворень, як-от шум, втрати пакетів і кодування.

З математичної точки зору він використовує кілька етапів перетворення сигналу. Один із них – перцептивне перетворення на частотний простір, після чого обчислюються об'єктивні метрики різниці між оригіналом та спотвореним сигналом. Формула, яка описує один із етапів, представлена як:

$$MOS_{PESQ} = \alpha * \|y_{ref} - y_{test}\| + \beta, \quad (1.2)$$

де y_{ref} – спектральне подання оригіналу сигналу;

y_{test} – спектральне подання спотвореного сигналу;

$\|y_{ref} - y_{test}\|$ – норма відмінностей між спектрами;

α та β – параметри моделі, які налаштовуються під час навчання.

Параметр α контролює вагу різниці між спектрами в фінальній оцінці, тобто визначає, наскільки важливі відмінності в спектральних компонентах для результату.

Параметр β додається для коригування фінальної оцінки MOS, щоб точно адаптувати результат до реального сприйняття людиною.

Формула описує оцінку якісті спотвореного сигналу, порівнюючи його з оригінальним. Основний акцент робиться на вимірюванні спектральних відмінностей між оригіналом та спотвореним сигналом. Чим менша різниця між y_{ref} та y_{test}, тим вищим буде MOS, вказуючи на хорошу якість спотвореного сигналу.

Більш вдосконалена версія PESQ це POLQA. Він не обмежується вузькосмуговими характеристиками телефонного зв’язку і охоплює як широкосмуговий 50 Гц – 14 кГц, так і суперширокосмуговий 20 Гц – 20 кГц діапазони. На відміну від PESQ, впроваджує динамічне вирівнювання часових шкал із адаптивним підстроюванням до затримок, завдяки чому навіть локальні затримки й джиттер не призводять до заниження оцінки. У спектральній області замість жорсткого фільтра нижніх та верхніх частот реалізовано гнучке згладжування спектру, яке дозволяє більш точно розпізнати вузькосмугові артефакти та зміни гармонік при компресії. Перцептивна модель POLQA збагачена ймовірнісною оцінкою фази й рівня шуму в критичних зонах сприйняття, особливо в області 2–5 кГц, тоді як PESQ оперує передусім з амплітудними відмінностями.

Удосконалюючи підходи PESQ і POLQA, VISQOL переорієнтовується

з простих моделей слухового сприйняття на розширену психоакустичну модель. VISQOL – це сучасна система оцінки якості звуку, яка була розроблена з метою надання більш точної та універсальної оцінки якості аудіо в порівнянні з традиційними методами. Замість застосування спрощених уявлень про слухову систему, вона використовує розширену психоакустичну модель, дозволяючи коректно оцінювати аудіосигнали зі складними типами спотворень. До таких спотворень належать шум, артефакти кодування, а також часові та частотні викривлення. Оцінювання в VISQOL починається з перетворення аудіосигналу у спектрограму за допомогою STFT для відображення структуру сигналу в часо-частотній області (рисунок 1.1). Спектрограми тестового й еталонного сигналів аналізуються за допомогою психоакустичних моделей, враховуючи частотну та амплітудну чутливість людського слуху. Остаточна метрика якості формується як зважена сума абсолютних розбіжностей між відповідними спектральними компонентами, де кожен ваговий коефіцієнт відображає перцептивну значущість відповідного частотного діапазону.

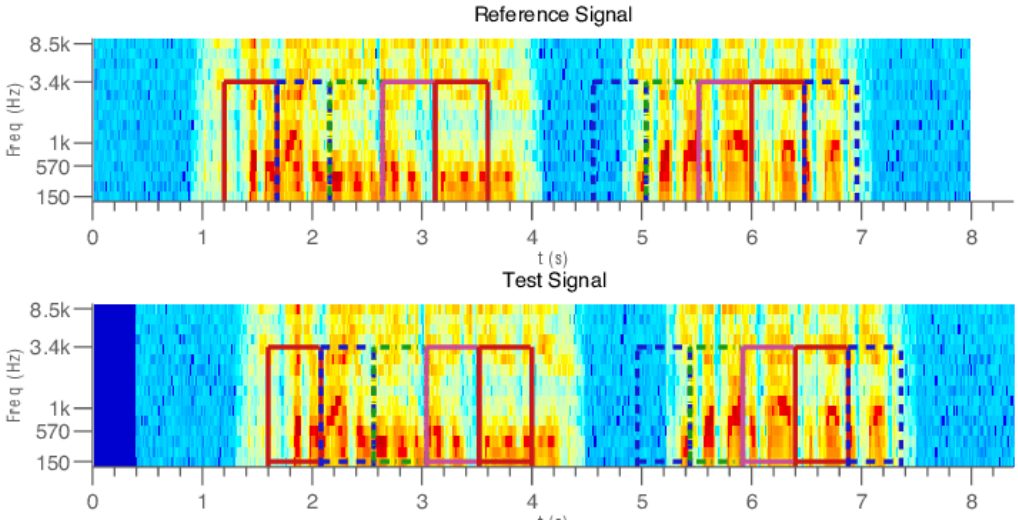

Рисунок 1.1 – Порівняльні спектрограми еталонного та тестового аудіо

Метрика SDR відображає співвідношення енергії «чистого» сигналу та енергії небажаних компонентів, які з’являються під час обробки або розділення аудіоканалів. По суті, вона кількісно вимірює, наскільки

відновлений чи виділений фрагмент відтворює амплітудно-спектральні властивості оригіналу. Обчислення полягає в порівнянні суми квадратів амплітуд корисного сигналу з сумою квадратів артефактів; чим менше «завад» – тим вищий показник. Високе значення SDR свідчить про точну реконструкцію або чітке розділення джерел, тоді як його зниження фіксує збільшення спотворень і втрат ключових аудіохарактеристик. Якщо відновлений сигнал має схожі амплітуди з вихідним, спотворень мало, тоді SDR покаже високе значення. А коли відновлений сигнал сильно спотворений, спотворень буде багато, то він буде низьким.

Істотним обмеженням метрики SDR є її чутливість до масштабування сигналу, призводячи до некоректної оцінки у випадках амплітудного зсуву між відновленим та еталонним сигналами. У відповідь на це було запропоновано модифікацію – SI-SDR, яка усуває залежність від рівня гучності шляхом нормалізації амплітуди. Фокусується він на структурній подібності сигналів для більш релевантної оцінки якості реконструкції.

В обробці аудіоданих обидва індикатори використовуються в завданнях розділення джерел, поліпшення мовної прозорості та придушення шуму. Починаючи з появи глибинних архітектур, SI-SDR здобула поширення як функція втрат у моделях на кшталт Conv-TasNet і SISDRNet. Нормалізація амплітуди дозволяє мережам ефективніше мінімізувати спотворення, забезпечуючи відновлення або сепарацію сигналів із відтворенням їхньої перцептивної цілісності незалежно від амплітудних відмінностей.

1.3 Застосування машинного навчання та нейромереж в аудіоаналізі

Нейронні мережі в аудіоаналізі вирізняються здатністю автоматично виявляти релевантні ознаки зі спектральних або часових представлень сигналу, усуваючи необхідність ручного добору параметрів. На відміну від традиційних алгоритмів цифрової обробки, які оперують заздалегідь заданими фільтрами та фіксованими спектральними метриками, глибинні

моделі навчаються виявляти складні патерни – від гармонічних співвідношень і динамічних змін до ритмічних структур і варіацій спектральної густини.

1.3.1 Рекурентні мережі в аудіоаналітиці: RNN, GRU та LSTM

Однією з перших спроб застосувати рекурентні нейромережі до поліфонічної музики було поєднання RNN із умовною Restricted Boltzmann Machin у роботі «Modeling Temporal Dependencies in High-Dimensional Sequences», для генерації поліфонічної музики на рівні MIDI-нот. Він став одним з перших, що намагався вловити залежності між акордами, послідовністю нот та їх гармонійною структурою. Демонструючи здатність відтворювати впізнавані стильові особливості композиторів, зберігаючи логіку розвитку музичної фрази, був проривом на той час.

Проте при застосуванні RNN до обробки високодеталізованого звукового сигналу виявилися обмеження, пов’язані із згасанням та вибухом градієнтів, вперше формально описані в статті «Learning long-term dependencies with gradient descent is difficult» [20]. У ній показано, що з ростом довжини оброблюваної послідовності похідні функції втрат експоненційно затухають або вибухають, що призводить до нерегулярного або відсутнього оновлення параметрів і нестабільності навчання. У результаті – параметри мережі або не оновлюються, або оновлюються неконтрольовано, що призводить до повної втрати стабільності під час навчання. Обробляти адекватно довгі музичні фрази або аудіофрагменти не була здатна зберігати взаємозв’язки на часовій відстані в кілька секунд або тисяч таймстепів.

У статті «Training and Analysing Deep Recurrent Neural Networks» автори продемонстрували оптимізовані функцій активації стандартна рекурентна мережа з глибиною понад 100 тимчасових кроків зазнає зниження точності: у задачі класифікації музичних подій її показники були

на 20–30 % нижчими порівняно з LSTM. Аналогічно, у роботі «Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling» доведено, що звичайні RNN не можуть стабільно відтворювати мел-спектрограми довжиною понад 2с без втрати фазової цілісності або зсуву амплітуди.

Неможливість стандартних RNN зберігати релевантну інформацію на великих часових відстанях спонукала до архітектурних змін у самому рекурентному блоці. Першою такою модифікацією стала архітектура Long Short-Term Memory, а пізніше – Gated Recurrent Unit. Обидві моделі вводять параметризовані "ворота" для гнучкого контролю потоку інформації через рекурентну клітину. З огляду на поширеність застосування підходів у задачах аудіоаналізу, доцільно порівняти їх за впливом на якість відновлення, швидкодію та адаптивність до музичних сигналів. У таблиці нижче наведено ключові переваги та обмеження кожного підходу з позиції їх застосування в контексті аудіоаналізу (таблиця 1.1).

Таблиця 1.1 – Технічні переваги та обмеження архітектур LSTM і GRU

Критерій	LSTM	GRU
1	2	3
Здатність зберігати довготривалі залежності	Має окремий "cell state", який передає інформацію незалежно від основного потоку – забезпечує високу стабільність пам’яті навіть на > 1000 кроків.	Простішу структуру із менш чітким розділенням пам’яті, що призводить до поступового згасання контексту.
Інтонаційна динаміка в генерації аудіо	Більш точно передає плавні зміни гучності й тембру, критично важливі для вокалу.	Добре справляється з ритмічною структурою, але менш стабільна при генерації вокальних фраз.

Продовження таблиці 1.1

1	2	3
Поводження з короткими повторюваними структурами	Має тенденцію до надлишкового накопичення контексту, що спотворює повторювані мотиви.	Добре фільтрує "локальні шаблони", завдяки агресивнішому оновленню пам’яті через reset gate.
Ресурсоємність виконання моделі в режимі реального часу	Високі затримки на forward-pass – не підходить для low-latency систем.	Використовується у реальних продуктах – працює в браузерах з <10 мс затримки.
Чутливість до зміщення фази в аудіосигналі	Нестабільна, особливо без доповнення згортками – створює флаттеринг.	Аналогічні проблеми, але швидше адаптується при коротких фрагментах.
Придатність до генеративних задач	Має схильність до згладжування спектра погано відновлює мікродеталі.	Швидше генерує вихід, але жертвує точністю відновлення.
Гнучкість для поєднання з іншими архітектурами	Вимагає складної інтеграції та глибшого налаштування таймінгів.	Простіше масштабується у гібридні архітектури.
Практичне застосування в open-source рішеннях	Використовувався в Deep Voice, WaveNet, але поступово витіснений attention-моделями.	Залишається актуальним у low-resource системах.

1.3.2 Згорткові архітектури для спектрального аналізу: CNN і U-Net

Потреба в паралельній обробці, гнучкій генерації спектральних компонентів і точному збереженні фазових характеристик сигналу вимагає

використання більш адаптованих архітектур. Архітектурою, яка найкраще відповідає цим вимогам, стала CNN. Згортки в CNN виконують роль частотно-часових фільтрів, що навчаються безпосередньо з даних, уникаючи заздалегідь визначених перетворень. Корисно при роботі з обертонами, фазовими накладеннями та спектральним перекриттям – компонентами, що не мають чіткої позиції, але є визначальними для музики. У роботі "Deep Learning for Audio Signal Processing" демонструється, що CNN-архітектури здатні успішно виявляти повторювані частотно-часові структури в умовах складної поліфонії, де класичні алгоритми спектрального фільтрування втрачають роздільну здатність.

Однак прості згорткові мережі мають суттєвий недолік – втрату дрібномасштабної інформації на глибших рівнях через багаторазове підсумовування та пулінг. Для аудіосигналів означає ризик втрати нюансів, що визначають тембр або мікродинаміку. Саме для збереження цих ознак була запропонована архітектура U-Net [8]. Вирізняється симетричною структурою енкодера та декодера із горизонтальними зв’язками між відповідними шарами (рисунок 1.2). Прямі зв’язки U-Net передають низькорівневі деталі під час декодування, мінімізуючи втрати й зберігаючи точність високочастотного спектру.

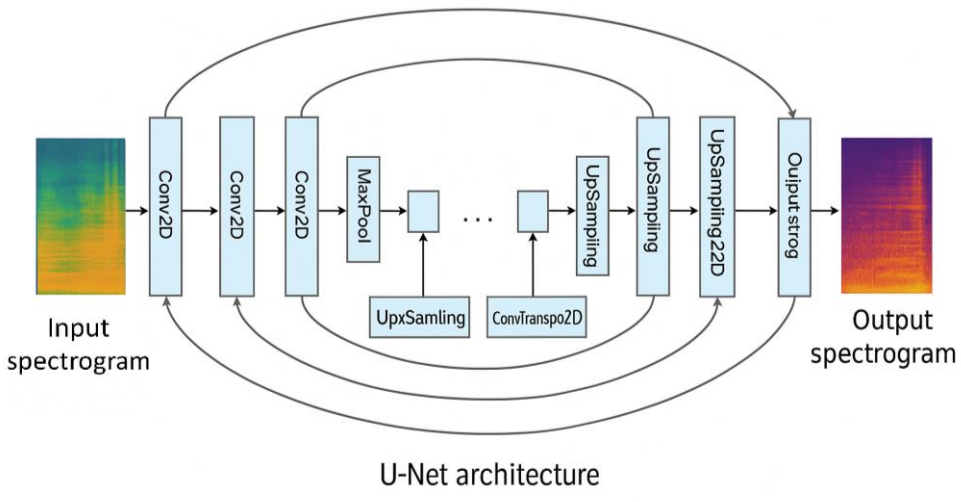

Рисунок 1.2 – Архітектура U-Net, адаптована для спектральної обробки аудіосигналів

1.3.3 Self-Attention і трансформери в моделюванні музичних сигналів

Механізм самоуваги (Self-attention) усуває структурну обмеженість U-Net [21], пов’язану з фіксованою областю прийняття та нездатністю моделі враховувати залежності між віддаленими фрагментами спектру. Незважаючи на наявність skip-з’єднань, інформаційний потік в U-Net формується на основі локальних обчислень, обмежених простором згорткового ядра. Кожен елемент латентного представлення залежить лише від певної кількості сусідніх фреймів у часі та частоті, що перешкоджає захопленню зв’язків, які не вкладаються в локальне вікно, але мають глобальну структурну функцію.

Self-attention дозволяє моделі одночасно аналізувати усі частини сигналу, встановлюючи між ними динамічні контекстні зв’язки (рисунок 1.3). Обчислення уваги ґрунтується на зіставленні кожного фрагмента спектру з усіма іншими, для гнучкого формування внутрішнього представлення сигналу без жорсткої прив’язки до локального вікна. У результаті модель має змогу відслідковувати темпоральні або спектральні залежності будь-якої довжини та щільності, формуючи взаємозв’язки, які відповідають логіці розвитку музичного матеріалу. Актуально у випадках, коли звукові події віддалені одна від одної, але мають спільні тембральні або ритмічні характеристики. У музичних творах зв’язки можуть реалізовуватись через повтори теми, варіації гармонічної послідовності чи синкоповані переходи, які втрачаються при фрагментарному аналізі. Self-attention адаптується до вхідного матеріалу, дозволяючи кожному фрагменту спектру впливати на обробку інших – незалежно від їхнього положення у часовій послідовності.

Трансформерна обробка не зводиться до локального зчитування шаблонів, а формує цілісне представлення сигналу, з урахуванням внутрішньої організації. У межах одного проходу модель формує карту ваг, яка відображає ступінь взаємозалежності між окремими елементами спектру. Дозволяючи зберігати логічну структуру композиції та координувати обробку її окремих частин відповідно до ролі у загальній акустичній картині.

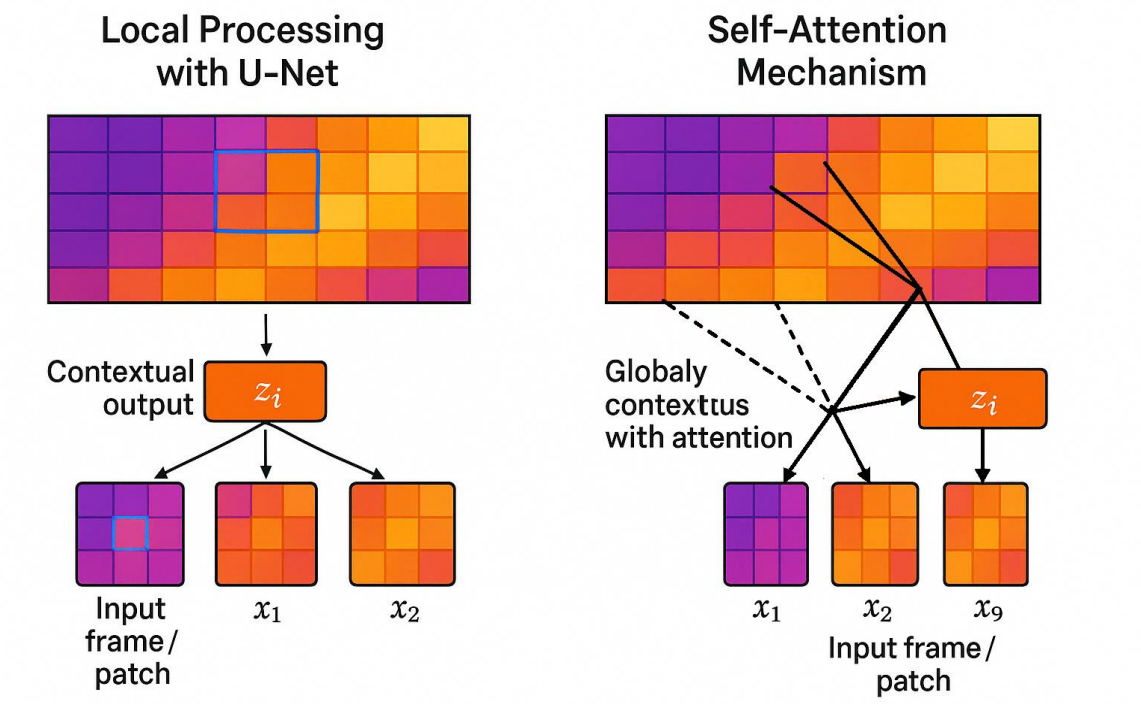

Рисунок 1.3 – Порівняння локальної та глобальної області взаємодії у моделях

1.4 Огляд існуючих досліджень та рішень у галузі оцінки якості звуку

У роботах, присвячених підвищенню аудіоякості за допомогою нейронних мереж, ключовим аспектом є методологія оцінювання результатів обробки. Механізми, що імітують слухове сприйняття, дедалі частіше витісняють класичні метрики, засновані на статистичних порівняннях сигналів.

Система DNSMOS, запропонована в рамках конкурсу Microsoft DNS Challenge, продемонструвала практичну ефективність для автоматичного прогнозування перцептивної оцінки якості. Модель побудована на основі спектральних ознак необробленого сигналу, працює без еталонного зразка та формує значення, що наближені до MOS-оцінки реальних слухачів. Показники кореляції з результатами прослуховування перевищують аналогічні характеристики традиційних алгоритмів. DNSMOS

використовується у тестуванні моделей FullSubNet, DCCRN, DeepFilterNet та інших систем, орієнтованих на шумозаглушення або покращення мови.

Архітектура MetricGAN реалізує альтернативний підхід: замість зовнішньої метрики оцінка якості вбудовується безпосередньо у навчальний цикл. У процесі оптимізації дискримінатор передбачає значення перцептивної метрики, а генератор модифікує сигнал з урахуванням цього прогнозу. Принцип орієнтується на бажаний результат у суб'єктивному сенсі, а не лише на точність відновлення. Він застосовується у його варіаціях – MetricGAN+, NoisyMetricGAN, Attention MetricGAN, демонструючи поліпшення перцептивної якості у складних акустичних умовах.

Окремо варто виділити архітектури, призначені для виконання в реальному часі. DeepFilterNet, зокрема, розроблена для розгортання на мікроконтролерах та DSP-системах. Складається з GRU-блоків і спектрального маскування з постфільтрацією, здатна адаптивно зменшувати шум без значних фазових спотворень. Якість роботи моделі оцінюється комбінацією метрик – SI-SDR, PESQ, DNSMOS та результатами слухових експериментів. Навіть у разі низької обчислювальної складності система досягає показників, порівнюваних із повнофункціональними рішеннями.

У контексті академічних і прикладних досліджень використовуються інструменти для автоматизації процесу вимірювання якості: бібліотеки speechmetrics, torchmetrics.audio, aud-eval, а також open-source реалізація visqol. Платформа VISQOL, розроблена Google, демонструє стабільну кореляцію з MOS-оцінками на музиці, що робить її придатною для використання у системах оцінки якості немовленнєвих аудіоданих.

Паралельно з метриками, які обчислюють відстань між сигналами, розвивається клас моделей-прогнозувачів якості. Серед них – MOSA-Net, UniEval-Audio, NoisyMOS, побудовані на базі трансформерів або каскадних згорткових мереж. Вони не потребують наявності референсу та працюють безпосередньо на аудіопотоці, формуючи оцінку на основі багатовимірного представлення спектру, фазової структури й тембральних характеристик.

Оцінка якості аудіо поступово переходить у домен високорівневого контекстного аналізу. Питання зводиться не до того, наскільки сигнал близький до оригіналу з технічної точки зору, а до того, як він сприймається у практичному середовищі. Вимагаючи комбінованих підходів, у яких інструментальні метрики поєднуються з нейромережевими предикторами, адаптованими під конкретні задачі – від очищення голосу в реальному часі до перцептивного відновлення музичних творів.

2 ТЕХНОЛОГІЧНІ ЗАСОБИ ТА СЕРЕДОВИЩЕ РЕАЛІЗАЦІЇ

2.1 Постановка задачі та визначення цільових характеристик аудіоякості

Оцінка якості аудіосигналів базується на кількісному аналізі параметрів, що описують співвідношення сигналу до шуму, динамічний та частотний діапазони, рівень нелінійних спотворень і стабільність стереоканалу. Вивчення фізичних властивостей сигналу дозволяє сформулювати критерії, за якими визначається придатність аудіоматеріалу до подальшої обробки або оцінювання.

2.1.1 Параметри фізичної якості аудіосигналів

Одним із базових параметрів, що визначає якість аудіосигналу, є співвідношення SNR. SNR – визначає якість аудіозапису та його придатність для різних застосувань. Його значення виражається у dB і є логарифмічним співвідношенням між рівнем фонового шуму, і рівнем корисного сигналу. Чим вище цей показник, тим чистішим і розбірливішим звучить аудіозапис. Однак важливо розібратися, як шум впливає на сприйняття. Шум за своєю природою буває різним і виявлятися у вигляді: білого, рожевого, червоного шуму, електромагнітних перешкод та цифрових артефактів [9]. Білий шум має рівномірну щільність спектру, що робить його поширеним у системах тестування. Рожевий шум характеризується рівномірною енергетикою на октаву, наближаючи його до сприйняття природних акустичних середовищ. Червоний шум має більшу енергію на низьких частотах, роблячи його важливим у дослідженнях слухового сприйняття. Візуальне подання цих типів шумів можна побачити на спектрограмі (рисунок 2.1).

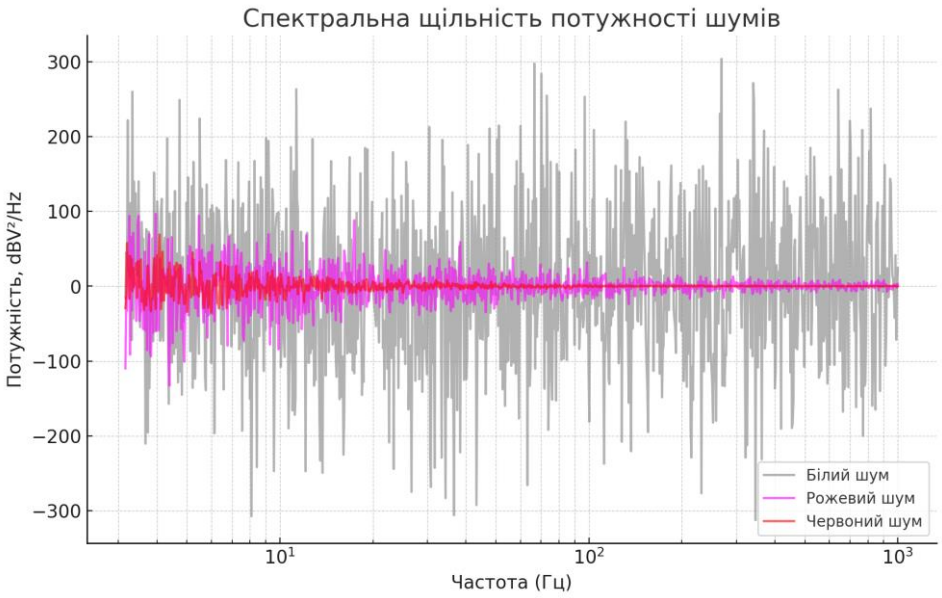

Рисунок 2.1 – Спектральна щільність потужності для різних типів шуму

Цифрові артефакти виникають при недостатній бітності сигналу, агресивному стисканні або неправильній обробці, що призводить до появи дискретизаційних шумів та спотворень. Електромагнітні перешкоди виявляються при недостатньому екрануванні аудіосистем та можуть вносити нестабільні шумові компоненти.

Динамічний діапазон – відображає різницю між найтихішим і найгучнішим звуком. Його величина виражається у dB та визначає можливість точної передачі звукових нюансів (рисунок 2.2). У цифровому аудіо діапазон обмежується розрядністю сигналу: для 16-бітного аудіо становить десь 96 dB, а для 24-бітного – до 144 dB. В аналогових системах діапазон залежить від рівня шуму та спотворень носія. У музичному продакшені широкий динамічний діапазон дозволяє природно передавати гучність вокалу та інструментів, уникаючи зайвої компресії, що робить запис плоским. В оркестровій музиці різниця між піаніссімо та фортисімо досягає близько 75-80 dB. Для рок- і поп-музики частіше використовується стиск, щоб досягти рівномірної гучності. У побутовому прослуховуванні теж відчувається: записи з вузьким діапазоном здаються стомлюючими через відсутність перепадів гучності. При використанні навушників та портативних

колонок компресія динаміки покращує розбірливість звуку в шумному середовищі, але позбавляє його глибини та природності. Слід зазначити, що сприйняття динамічного діапазону тісно пов'язане із психоакустичними особливостями людського слуху [18]. Наш слуховий апарат здатний сприймати звуки у діапазоні приблизно 120 dB, але має нелінійну характеристику чутливості. Означає, що люди більш сприйнятливі до змін гучності середніх і низьких рівнях сигналу і менш чутливі до таких змін на високих рівнях.

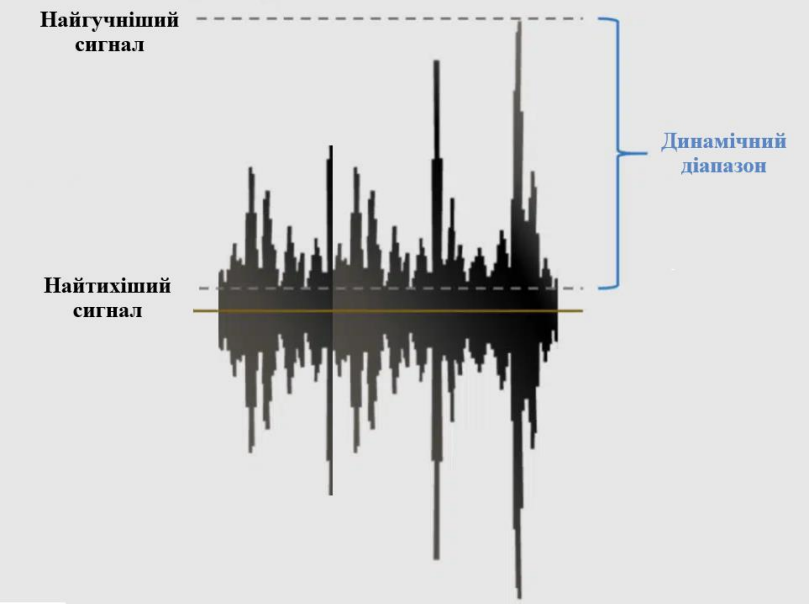

Рисунок 2.2 – Візуальне подання динамічного діапазону аудіосигналу

Частотний діапазон – визначає, які частоти аудіосистема або запис може відтворюватися, а також наскільки рівномірно це відбувається. Виражається як діапазон від найнижчої до найвищої частоти, яку система здатна адекватно передати. Більшість аудіосистем цей діапазон лежить у межах від 20 Гц до 20 кГц, що збігається з діапазоном слухового сприйняття людини. Вузький частотний діапазон позбавляє важливої інформації: недостатнє відтворення низьких частот робить баси «плоскими», а слабка виразність верхів – звук каламутним і непрозорим. У той же час занадто широкий діапазон створює труднощі у відтворенні сигналів з низькими і

високими частотами, якщо система не здатна точно обробити екстремальні значення. Не менш критичною характеристикою частотного діапазону виступає його рівномірність. Аудіосистема з рівним відгуком забезпечить, що всі частотні діапазони будуть передані без викривлень і спотворень. Коли частотний відгук системи не рівномірний виявляються "провали" або "піки" на окремих частотах (рисунок 2.3). При прослуховуванні запису здається, що низькі частоти є надмірними, або навпаки, що високі частоти сильно пригнічені.

Рисунок 2.3 – Амплітудно-частотна характеристика аудіосистеми

Гармонічні спотворення – це коли звук набуває «присмаку», якого не було в оригіналі. Вимірюється ступінем нелінійних спотворень, що виникають в аудіосистемах під час обробки або передачі сигналу. Аудіопристрій або система не є ідеально лінійною, оригінальний сигнал перетворюється таким чином, що в його спектрі з'являються додаткові гармоніки – частоти, кратні основним частотам сигналу. Гармоніки створюють спотворення, які можуть істотно змінити сприйняття звуку. Основна роль тому, що він дозволяє оцінити, наскільки чисто аудіосистема відтворює вихідний сигнал, коли система або пристрій мають високі

гармонічні спотворення, призводить до появи небажаних призвуків чи артефактів, які можуть погіршити сприйняття. Якщо підсилювач додає гармоніки на частоті, що у вдвічі перевищує основну, зайві частоти можуть накластися на інші елементи твору, створюючи ненатуральне звучання, не всі гармонічні спотворення однаково сприймаються людиною. У деяких випадках такі спотворення можуть бути навіть сприйняті як бажані, особливо в контексті музичних інструментів або певних жанрах музики.

Перехресні перешкоди – це ніби два сусіди розмовляли через стіну, але через погану звукоізоляцію їх голоси частково змішувалися. У результаті чується не тільки того, з ким говориш, але й приглушені фрази іншого, що заважають сприйняттю. Приводить до того, що звуки, призначені тільки для одного каналу, непомітно «витікають» в інший, руйнуючи чіткість стереоефекту та глибину звукової сцени. Перехресні перешкоди відбуваються через різні фактори: недосконаліть у компонентах аудіосистеми, у підсилювачах або проводах, які призводять до неконтрольованого «перенесення» сигналу між каналами. В ідеалі кожен канал у стерео або багатоканальній системі повинен передавати свій сигнал без будь-якого впливу на інших, щоб зберегти чіткість стереоподілу. Однак, на практиці завжди існує деяка кількість перехресних перешкод, і їх величина може бути дуже різноманітною в залежності від якості використовуваної техніки. У дешевих підсилювачах або аудіообладнанні з недостатньою ізоляцією, сигнал одного каналу просочується в сусідній. Призводить до того, що звуки, які мають бути відтворені лише одним каналом, починають чутись в іншому каналі. Такий ефект порушує стереорозподіл та погіршує сприйняття просторовості звуку (рисунок 2.4). Коли перехресні перешкоди збільшуються, зникає відчуття звукової сцени, тобто уявлення про те, що звуки виходять із певних джерел, розташованих у просторі.

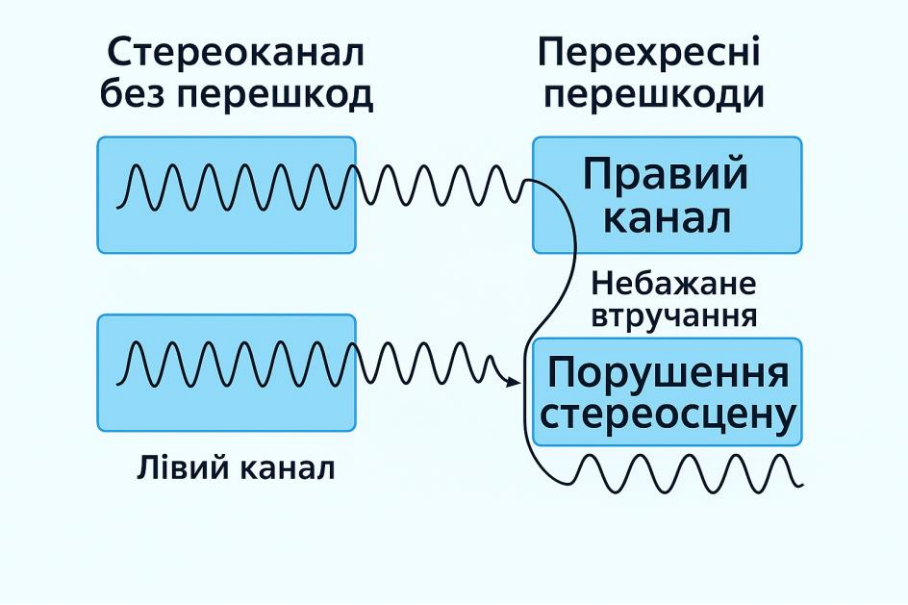

Рисунок 2.4 – Схема перехресних перешкод у стереоаудіосистемі

2.1.2 Постановка задачі підвищення якості аудіосигналів

Порушення фізичних характеристик аудіосигналів, серед яких співвідношення сигналу до шуму, спектральна повнота, динамічний діапазон і просторове розділення каналів (див. підпункт 2.1.1), призводить до істотного зниження якості відтворення як у технічних, так і у психоакустичних вимірах. Наявність шумових домішок, обмеження амплітудного діапазону, деформація спектральної структури та збільшення рівня перехресних перешкод обмежує точність відтворення деталей запису [11], спотворює природність звучання та погіршує суб'єктивне сприйняття аудіоматеріалу.

Задача підвищення якості формулюється як процес відновлення фізичних параметрів сигналу, порушених внаслідок деґрадацій. Мета полягає у зменшенні рівня шумів, компенсації втрат динаміки, вирівнюванні частотної характеристики та мінімізації гармонічних спотворень без внесення додаткових артефактів. Пріоритетом є збереження природної структури сигналу, що включає спектральні залежності та просторову розподіленість джерел звуку.

Підхід реалізується у межах парадигми навчання з учителем, де вхідними даними слугують аудіофрагменти зі штучно створеними спотвореннями, а вихідними – відповідні референсні зразки без дефектів. Модель має наближати вихідний сигнал до еталону з урахуванням оптимізації за декількома критеріями якості.

Для вирішення задачі відновлення аудіоякості необхідно використовувати представлення сигналу, забезпечуючи максимальне збереження його структурних особливостей при обробці в нейронних мережах. За результатами досліджень «Spectral Representations for Enhanced Audio Modeling», застосування спектральних перетворень, зокрема спектрограм та мел-спектрограм, суттєво підвищує ефективність моделювання в порівнянні з обробкою сирих сигналів у часовій області.

Переваги спектрального представлення пояснюються тим, що спектрограми забезпечують явне відображення частотних та енергетичних характеристик сигналу у компактній формі для точного відновлення втрачених компонентів. Згідно з висновками «Deep Learning for Audio Signal Processing» спектральні ознаки в аудіозавданнях дозволяють досягти вищої точності при реконструкції, ніж хвильова форма, особливо у випадках, коли сигнал зазнає складних типів спотворень.

2.2 Представлення аудіосигналу для нейронної обробки

Успіх алгоритмів підвищення аудіокацтва визначається архітектурою та способом «подання» звукових даних до мережі. Оптимальна вистава фіксує критичні спектро-темпоральні деталі, залишається обчислювально прийнятною і допускає градієнтне навчання. При виборі формату враховують три групи факторів:

- збереження спектро-темпоральних деталей (інакше спотворення залишаться невиявленими);
- розмір та топологія тензора, визначає пам'ять та глибину моделі;

- інвертованість чи пряму відповідність психоакустичним шкалам, що полегшує кількісну оцінку поліпшень.

На практиці, більшість існуючих підходів до покращення аудіоякості представляють вхідний сигнал у спектральному або спектро-темпоральному вигляді. Розглянемо докладніше основні формати та його характеристики.

2.2.1 Спектральні уявлення

Спектрограми – це двовимірні масиви, які отримують після короткочасного перетворення Фур'є, де вертикальна вісь відповідає частоті, а горизонтальна – часу. Серед спектрограми [15] виділяють кілька поширених видів (рисунок 2.5).

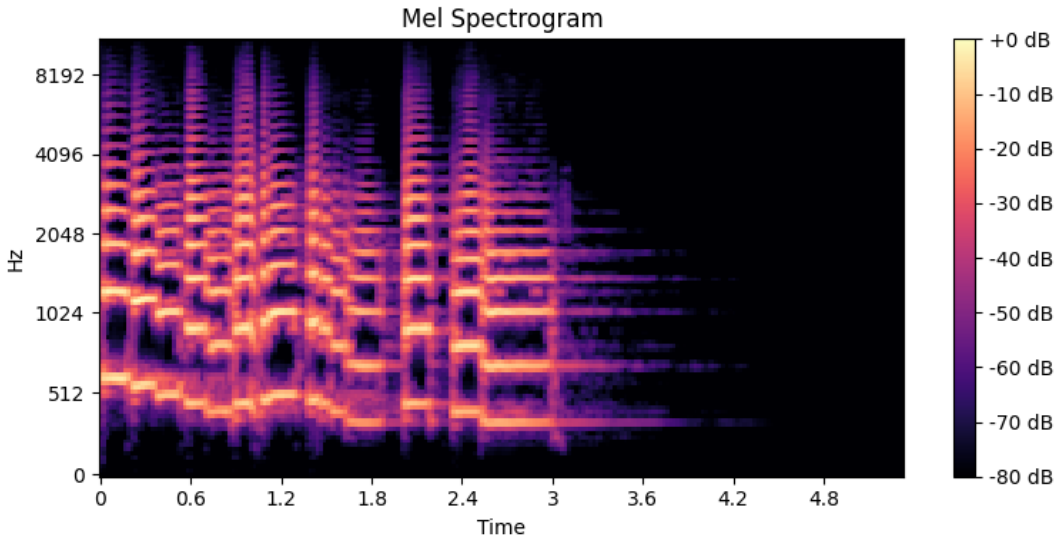

Рисунок 2.5 – Мел-спектрограма аудіосигналу

STFT є базовим та найбільш універсальним форматом, що формує комплекснозначну спектральну матрицю, що описується виразом:

$$X(t,f) = \sum_{n=-\infty}^{\infty} x[n]\omega[n-tH]e^{-j2\pi fn/N}, \qquad (2.1)$$

де x[n] – вхідний сигнал;

ω[n] – віконна функція (зазвичай Ханна або Хеммінга);

H – крок вікна;

N – розмір вікна FFT.

Виходячи з цього, комплексна спектрограма X(t,f) містить повну інформацію про тимчасові та спектральні характеристики сигналу, проте вона надмірна за розмірністю і не враховує психоакустичні особливості сприйняття людиною звукових частот. Для усунення даних недоліків використовують логарифмічну Mel-спектрограму, засновану на нелінійному перетворенні частотної осі за Mel-шкалою і задається формулою:

$$f_{mel} = 2595 * \log_{10}(1 + \frac{f_{Hz}}{700}), \quad (2.2)$$

де f_{mel} – частота, виражена в Mel-шкалі;

f_{Hz} – частота в герцах;

2595 – є масштабуючим коефіцієнтом, що забезпечує відповідність сприйняття тональних змін логарифмічного характеру слуху;

700 – відповідає частоті перегину, після якої слух переходить від лінійного сприйняття частоти до логарифмічного;

Log-Mel спектрограма формується у два етапи.

1 Етап. Застосування банку Mel-фільтрів до амплітудного спектру, отриманого STFT. Вони є набором трикутних фільтрів, що мають рівномірно розподілені центри частот за шкалою. Застосування Mel-фільтробанку зводиться до операції множення амплітудного спектру на ваги фільтрів:

$$\mathrm{M}(n,m) = \sum_{k} |X(n,k)|^2 * W_k\ (k), \quad (2.3)$$

де $W_k\ (k)$ – коефіцієнти Mel-фільтробанку;

m – номер Mel-частотного каналу.

2 Етап. Переклад у логарифмічний масштаб для посилення динамічного розмаїття та нормалізації розподілу енергії:

$$\mathrm{M}\log(n,m) = 10 * \log_{10}(M(n,m)+ \in), \tag{2.4}$$

$\in$ – невелика постійна величина, що запобігає помилкам при обчисленні логарифму нуля.

В результаті описаних перетворень формується логарифмічна Mel-спектрограма, що є ефективною формою вхідних даних для нейронних моделей за рахунок логарифмічного масштабування частотної осі та агрегування спектральної енергії в обмежену кількість частотних каналів. Знижується розмірність вихідного спектрального уявлення, виділяється найбільш значущі спектро-темпоральні характеристики сигналу, забезпечуючи рівномірний частотний дозвіл та оптимізувати обчислювальні ресурси.

2.2.2 Розмірність та форма тензора

Спектрограма лог-Mel, що передається в мережу, є тривимірним масивом B × F × T, де B – обсяг пакету, F – число частотних каналів, T – кількість тимчасових кадрів. При типових налаштуваннях (F ≈ 128, крок вікна 10 мс, частота дискретизації 16 кГц) одна секунда музики конвертується в матрицю порядку $1{,}3 \times 10^4$ елементів, що більш ніж удвічі менше вихідного об'єму відліків і вже впорядковано по частоті. Лінійні за розміром входу згортки збільшують обчислювальну роботу пропорційно добутку F×T, тому подвоєння частотного дозволу 256 замість 128 каналів, призводить до еквівалентного зростання операцій, а отже, і до зниження допустимого розміру B. У багаторівневих U-Net-схемах, що працюють у часовій області, ускладнюється вимогою, щоб кожна з осей ділилася на 2^k,

інакше доводиться вводити додатковий паддинг, що збільшує проміжні тензори і породжує прикордонні артефакти, що чітко виявилося у Wave-U-Net. Для трансформерів ситуація ускладнюється квадратичною залежністю пам'яті self-attention [19] від $N = F \times T$. Аналіз обчислювальної складності доводить, що без спеціальних розріджених або проекційних схем $O(N^2)$ є нижньою межею, а тому при $T \approx 100$ і $F \approx 128$ повний механізм уваги вже вимагає декількох гігабайт на один приклад. Метод Patchout вирішує проблему, випадково відкидаючи частину патчів спектрограми на етапі навчання і тим самим знижуючи N та витрату GPU-пам'яті на порядок без втрати точності. Інший шлях оптимізації – адаптивне ущільнення часової осі. Диференційовано зливаючи малозначущі кадри, заощаджуємо не менше 20-25 % операцій при збереженні або навіть прирості якості класифікації [13], оскільки мережа вчиться ігнорувати надлишкові фрагменти та концентруватися на інформативних подіях. У практичних сценаріях рекомендується обмежувати частотну розмірність діапазоном 128–256 каналів та вибирати крок вікна 10 мс як компроміс між детальністю та обсягом пам'яті. При необхідності тоншого тимчасового аналізу тимчасову дискретизацію знижувати до 5 мс доцільно лише за наявності ≥ 16 ГБ відеопам'яті. При всій значущості оптимального рангу тензора вирішальна перевага спектрограм проявляється, коли їх коефіцієнти допускають впевнене відновлення тимчасової форми і одночасно розподіляють енергію за шкалою, що близька до суб'єктивних критичним смугами слуху.

2.2.3 Інвертованість чи пряме відповідність психоакустичним шкалам

Інвертованість спектрального домену, що вибирається, задає верхню межу досяжної якості синтезу: повнофазна STFT утворює надмірну рамку і відновлюється точно, тоді як відсутність фази вимагає ітеративного алгоритму Гріффіна - Ліма. Його «швидка» модифікація скорочує кількість ітерацій та збільшує відношення SNR при реконструкції. Лог-Mel-подання

зменшує розмірність шляхом підсумовування сусідніх ліній, що спрощує навчання, але робить зворотний перехід до наближених: відновлення досягають нейронні вокодери або псевдозворотні фільтри, які неминуче вносять частотну похибку (рисунок 2.6).

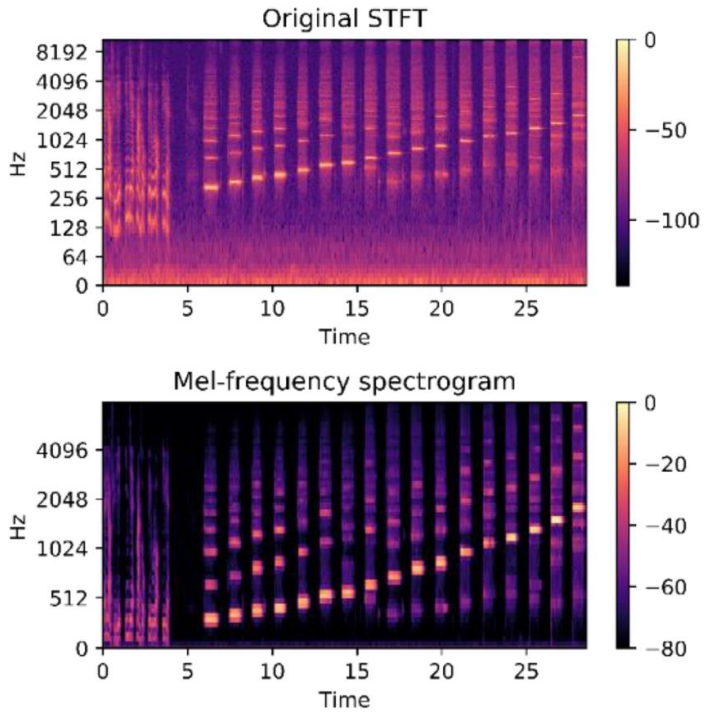

Рисунок 2.6 – Порівняння лінійної STFT-спектрограми та логарифмічної Mel-спектрограми

Для підвищення перцептивної релевантності фільтробанку підлаштовують під критичні смуги слуху: шкала Bark ділить діапазон на 24 зони з шириною, яка логарифмічно зростає після 500 Гц, відображаючи нерівномірну частотну вибірковість базилярної мембрани. Еквівалентна прямокутна ширина ERB задається емпіричною формулою Glasberg - Moore і дає безперервну апроксимацію смуг пропускання слухових фільтрів, дозволяючи будувати гамматонові банки, які пов'язують спектральну енергію з гучністю, що сприймається. Формується компроміс: суворо оборотні домени гарантують точний синтез за високих обчислювальних витрат, тоді як компактні психоакустичні шкали економлять ресурси ціною часткової необоротності.

2.3 Життєвий цикл набору даних

Безумовна точність та відтворюваність обчислювальних експериментів з нейромережевими моделями безпосередньо залежить від чітко організованого процесу підготовки даних. Помилки на цьому рівні призводять до спотворення навчальної вибірки, усунення метрик та некоректних висновків про ефективність архітектурних рішень. Тому етап роботи з даними вимагає не менш суворого формалізму, ніж проектування моделі або вибір функції втрат.

Структурування життєвого циклу у вигляді семи послідовних етапів узгоджується із загальноприйнятими підходами до управління даними в машинному навчанні. Кожен етап виконує конкретну функцію в ланцюжку підготовки, які сукупність охоплює весь процес – від відбору вихідних аудіофайлів до підсумкової валідації вибірки (рисунок 2.7).

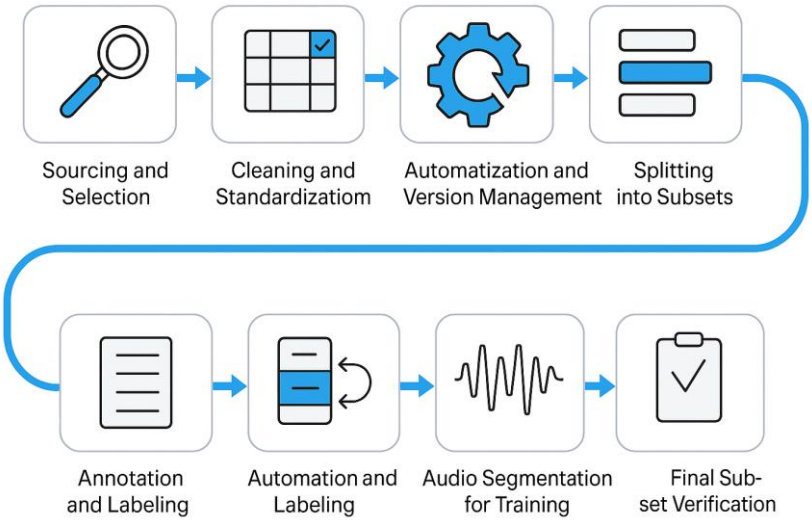

Рисунок 2.7 – Життєвий цикл підготовки аудіоданих для навчання

Початковий етап являє собою збір, фільтрацію та попередню верифікацію музичного матеріалу. Формування датасету починалося з відбору музичного матеріалу, що відповідає встановленим технічним та

змістовним критеріям: формат без стиснення WAV та FLAC, частота дискретизації не менше 44.1кГц, глибина – 16 біт, відсутність кліпінгу, рівномірна гучність, різноманітність жанрів та типів акустичних. Для підвищення стійкості моделі враховувалися як студійні, і «живі» записи. Аудіофайли, які відповідають вимогам, виключалися до етапу попередньої обробки.

Далі виконувались операції очищення та стандартизація. Усі аудіозаписи наводилися до єдиного формату: WAV PCM 16-bit, 44.1кГц, монофонічний канал. Проводилася уніфікація за тривалістю, нормалізація гучності, видалення тиші, усунення битих або порожніх фрагментів. Використання одного стандарту формату знижувало ризик помилок при завантаженні та обробці на стороні моделі.

До кожного аудіофайлу додавалися описові характеристики: жанр, спосіб запису, рівень шуму, динаміка, тривалість. Додатково обчислювалися аудіофічі: темп, спектральна щільність, частотний розкид, спектральна рівномірність. Розмітка зберігалася в структурованих таблицях з фіксованими ID, застосовувалася при стратифікації та подальшому аналізі результатів.

Усі операції з підготовки даних формалізувалися через систему DVC, де кожен крок описувався як декларативного пайплайна. Будь-які зміни вхідних даних або параметрів обробки автоматично фіксувалися. Кожному датасету відповідала унікальна контрольна сума, пов'язана з версією коду та конфігурацією моделі. Аудіокорпус ділився на три частини, що не перетинаються: навчальну, валідаційну та тестову. Використовувалася стратифікація за жанровою приналежністю, шумовою структурою, тривалістю та іншими ознаками. Усі підмножини зберігалися у вигляді окремої версії з повним описом розподілів та контрольних параметрів. Виключалися повтори та випадкові перетину.

Аудіозаписи нарізалися на сегменти фіксованої довжини зазвичай від 10 до 20 секунд. Застосовувалася нарізка з перекриттям, щоб підвищити

густину навчальних прикладів. Кожному сегменту присвоювався унікальний ідентифікатор, відбувалася повторна нормалізація гучності та перевірка амплітудних обмежень. Сегменти успадковували метадані батьківського файлу, що забезпечувало узгодженість на рівні батчів під час навчання моделі.

Вся вибірка перевірялася щодо відповідності формату, цілісності сегментів, коректності ID, правильної прив'язки до анотацій. Порівнювалися контрольні суми, генерувалася зведена таблиця з параметрами кожної підвиборки, описом структури, числовими характеристиками та версійними ідентифікаторами.

2.4 Вибір інструментів для реалізації (Python, PyTorch, Librosa, Torchaudio)

Як базова мова була Python, оскільки вона концентрує екосистему сучасних бібліотек для машинного навчання та цифрової обробки сигналів. Підтверджується даними статистичних досліджень сайту HiTech Service, який спеціалізується на аналітиці та IT-консалтингу, де Python стабільно займає першу позицію за популярністю серед інструментів для створення моделей нейронних мереж та аналізу даних (рисунок 2.8). Згідно з щорічним звітом GitHub Octoverse за 2024 рік, Python є найбільш поширеним інструментом розробки в галузі штучного інтелекту, аналізу даних та обробки сигналів, перевершуючи за кількістю активних репозиторіїв найближчі альтернативи JavaScript та Java. Подібні висновки підтверджуються і аналітикою IEEE Spectrum, де Python також займає лідируючі позиції щодо популярності та застосовуваності у наукових та інженерних завданнях.

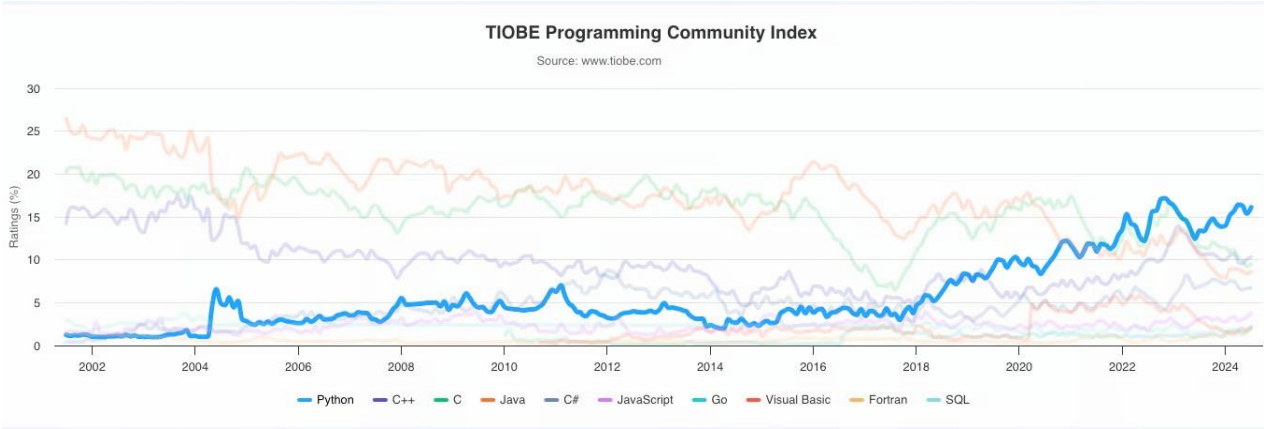

Рисунок 2.8 – Рейтинг найпопулярніших мов програмування в галузі машинного навчання та штучного інтелекту у 2025 році

Для побудови та навчання нейронних моделей обрано фреймворк глибокого навчання PyTorch, розроблений дослідницькою групою компанії Meta. PyTorch відрізняється динамічною побудовою графа обчислень, що забезпечує можливість зміни архітектури нейронної мережі безпосередньо в процесі виконання коду. Корисно при розробці та налагодженні рекурентних архітектур RNN, LSTM, трансформерів та каскадних мереж, які вимагають адаптації структури під конкретні експериментальні умови та варіації спектрального уявлення. Вбудовані засоби автоматичної диференціації, GPU-прискорення з використанням CUDA та cuDNN та інструмент профілювання ресурсів істотно підвищують продуктивність обчислень та дозволяють ефективно проводити експерименти з більшими обсягами даних.

Бібліотека NumPy була застосована для виконання чисельно ефективних операцій попередньої обробки та стандартизації аудіоданих. NumPy має високу швидкість за рахунок векторизації математичних операцій і низькорівневої оптимізації, завдяки чому скорочується час розрахунку спектрограм, нормалізації рівнів гучності, обчислення контрольних сум та інших базових задач. Підтримка багатовимірних масивів дозволяє ефективно виконувати масові перетворення сигналів та формувати стандартизовані вхідні тензори для нейромережевої моделі.

Для управління метаданими та підготовки анотацій використовувалася

бібліотека Pandas. Вона надає оптимізовані структури даних, здатні швидко завантажувати, фільтрувати та агрегувати таблиці з інформацією про сегменти аудіо, їх мітки якості та інші параметри. Оперативно виключати із навчальних вибірок сегменти з недостатньою тривалістю або низькою якістю, контролювати цілісність даних та виконувати агрегацію метрик, таких як середнє значення та дисперсія параметрів у рамках підвиборок [16]. Робота з тимчасовими мітками Pandas також дозволила точно синхронізувати сегменти аудіозаписів і відповідні їм анотації.

Візуалізація даних та результатів експериментів здійснювалася за допомогою Matplotlib. Дана бібліотека дозволила будувати наочні спектрограми до та після обробки нейронною мережею, графіки втрат та метрик у процесі навчання, забезпечуючи якісний та кількісний контроль за ефективністю придушення шумів та покращення аудіо. Підтримка різних форматів візуалізації та інтеграція з інтерактивними середовищами, такими як Jupyter Notebooks, додатково підвищила зручність та прозорість аналізу результатів експериментів.

3 РОЗРОБКА МОДЕЛІ ОЦІНКИ АУДІОЯКОСТІ МУЗИЧНИХ ТВОРІВ

3.1 Архітектурні передумови та побудова багатоетапної моделі оцінки аудіоякості

3.1.1 Аналіз архітектури DeepFilterNet2

Модель DeepFilterNet2 [12] є двоетапною нейромережевою архітектурою, призначеною для реалізації завдань шумоподавлення в реальному часі з фокусом на мовні аудіосигнали. Конструкція моделі орієнтована на низькорівневу обчислювальну складність та роботу в умовах обмежених апаратних ресурсів.

Обробка вхідного сигналу організована у два послідовні етапи. На першому етапі виконується спектральне перетворення сигналу із застосуванням фільтраційного банку, побудованого за шкалою ERB, що дозволяє отримати частотно-зважене уявлення, наближене до сприйняття людиною. Перетворення забезпечує спектральне декодування аудіопотоку в поданні, більш стійкому до коливань в області високих частот і допускає стиск спектру без значних втрат ключових мовних ознак. Другий етап включає нейромережевий модуль фільтрації, побудований на основі згорткових блоків з локальною рецептивною областю. Архітектура фільтра охоплює динаміку періодичних компонентів сигналу та забезпечує їх виділення у спектральній області. Завданням фільтра є придушення адитивного шуму за збереження гармонійної структури промови. Фільтрація виконується в комплексній формі впливає на амплітудні характеристики та фазову структуру сигналу.

Модель навчається з використанням комбінації спектральних втрат та перцептивних метрик, спрямованих на оптимізацію відновлення розбірливих та акустично несуперечливих мовних сигналів. Особливістю DeepFilterNet2 є

використання сегментованої обробки: вхідний сигнал розбивається на короткі часові кадри, дозволяючи забезпечити передбачувану латентність та стабільне навантаження при виході. Відсутність рекурентних або трансформерних компонентів робить архітектуру стійкою до затримок, але обмежує обсяг тимчасового контексту, що обробляється.

Конфігурація DeepFilterNet2 відображає підхід, орієнтований на збалансоване співвідношення між якістю придушення шуму та витратами на обробку. Основні рішення: використання ERB-подання, згорткової фільтрації та статичної двоетапної схеми забезпечують високу ефективність в умовах акустичних спотворень, типових для завдань мовної комунікації.

3.1.2 Обмеження двоступінчастої архітектури

У результаті попереднього аналізу моделі DeepFilterNet2 (див. п. 3.1.1) були ідентифіковані низка архітектурних та функціональних особливостей, які обумовлюють її ефективність у задачах обробки мовних сигналів [10]. Водночас ці ж рішення створюють обмеження при застосуванні до ширшого класу аудіоданих, зокрема музичних або змішаних акустичних сигналів.

Модель реалізована як статичний конвеєр з двома фіксованими етапами обробки, без урахування змін складності вхідного сигналу. Обробка здійснюється на коротких кадрах тривалістю до 5 секунд, що обмежує її здатність моделювати довготривалі залежності. Відсутність глобальних механізмів уваги, а також орієнтація на локальні ознаки спектру знижує її адаптивність до складних або нестандартних акустичних сценаріїв.

З метою деталізації виявлених відмінностей між DeepFilterNet2 та реалізованою багатоступеневою архітектурою у таблиці (таблиця 3.1) наведено порівняння за низкою технічно значущих критеріїв, що охоплюють структуру, функціональні можливості, адаптивність та стійкість моделей до артефактів.

Таблиця 3.1 – Порівняння архітектурних характеристик моделей

Критерій порівняння	DeepFilterNet2	Запропонована модель
1	2	3
Тип організаційної структури моделі	Статична двоетапна модель з фіксованою логікою проходження	Багатоетапна адаптивна модель з умовною маршрутизацією
Кількість функціонально незалежних етапів	Два послідовно з'єднані етапи	Чотири спеціалізовані етапи з незалежною функціональністю
Акустична спрямованість архітектури	Оптимізована переважно під вокальну мову	Універсальний підхід: адаптація до музики, мови, шумових міксів
Спроможність обробки немовних (музичних) сигналів	Складність масштабування та відсутність спеціалізованих підходів	Підтримка немовних аудіосигналів на архітектурному рівні
Ефективність при обробці довгострокових залежностей	Обмежений часовий контекст (5–10 секунд)	Розширене охоплення часового контексту через модульну обробку
Наявність механізмів глобальної уваги або трансформерів	Відсутні глобальні модулі уваги	Вбудовані елементи уваги та механізм маршрутизації
Підтримка деревербераці та придушення пізніх відлунь	Низька ефективність при наявності реверберацій компонентів	Підтримка приглушення реверберацій завдяки глибокій структурі
Стабільність при обробці складних акустичних сцен	Підвищений ризик спотворення складних музичних сегментів	Стійкість до акустичної складності та насиченості тембру

Продовження таблиці 3.1

1	2	3
Наявність адаптивного управління глибиною обробки	Відсутній динамічний механізм адаптації обробки	Наявність оцінки сигналу та умовної активації етапів обробки
Стійкість до генерації спектральних артефактів	Імовірність появи щебетання, фазових зсувів	Зменшення ймовірності артефактів завдяки поетапній фільтрації
Баланс між обчислювальною ефективністю та якістю	Сильна оптимізація під real-time, але за рахунок повноти відновлення	Збалансована архітектура зі збереженням якості та адаптивності

3.1.3 Обґрунтування багатоступеневої структури

Зіткнувшись з обмеженнями традиційних двоетапних рішень, стає очевидною необхідність гнучкішої стратегії обробки аудіо. Замість спроби вирішити всі завдання у межах однієї чи двох нейромереж, пропонується розділити процес кілька послідовних модулів, кожен із яких вирішує суворо певну задачу. У дослідженні "Glance and Gaze: Inferring Action-aware Points for One-Stage Human-Object Interaction Detection" автори представили двоетапний підхід, де перший етап, що називається «погляд», швидко видаляє грубий шум, тоді як другий етап, «погляд», фокусується на тонкій спектральній реконструкції для подальшого покращення вихідних даних. Їх дослідження показало, що поділ завдання на два спеціалізовані модулі – дало чудові результати у порівнянні з одноетапними моделями. Він привів до більш високих показників PESQ та STOI, продемонструвавши, як кілька етапів можуть допомогти покращити загальну якість та розбірливість завдань щодо покращення мови чи звуку.

Аналогічно, у роботі "Deep Multistage Multi-Task Learning for Quality

Prediction of Multistage Manufacturing Systems", автори запропонували багатоступінчасту систему покращення мови, включаючи шари внутрішньої уваги для більш точного покращення та динамічного регулювання глибини. Їхня модель продемонструвала, що поділ завдання поліпшення на кілька етапів, кожен з яких фокусується на різних аспектах сигналу, дає кращу продуктивність, ніж використання однієї монолітної мережі. Дослідження показало перевагу використання спеціалізованих етапів для покращення розбірливості мови та зниження залишкового шуму.

Враховуючи вищезгадані результати та обмеження монолітних нейромережевих рішень при роботі з реальними даними, в рамках роботи була реалізована архітектура, що складається з чотирьох спеціалізованих етапів. Кожен із них спроектований з урахуванням виявлених технічних закономірностей: початкова стадія відповідає за грубе придушення шумів, друга – за структурне уточнення з використанням уваги, третя – за зменшення артефактів через GAN, четверта – за адаптацію глибини обробки через систему зворотний зв'язок. Багатоетапна реалізація керує обробкою й маршрутизує сигнал залежно від стану.

3.2 Відбір музичних даних з урахуванням архітектурних вимог

Застосовувалися три категорії аудіосигналів: музичні твори, мовні сегменти та шумові профілі. Для кожної категорії виконувались завдання щодо навчання, валідації специфічних функцій моделі та її стійкості до різних спотворень.

3.2.1 Формування музичної підмножини корпусу

Для формування музичної частини корпусу використовувалися відкриті датасети GTZAN, MagnaTagATune та Free Music Archive, що надають доступ до ліцензованих музичних фрагментів у широкому жанровому діапазоні.

Аудіофайли охоплювали академічну та інструментальну музику, джаз, електронні композиції, акустичні записи та рок. Критерії відбору включали аналіз трьох спектрально-часових ознак:

- спектральна густина – оцінювалася за допомогою середнього значення спектральної енергії по частотних смугах, виділяючи насичені гармонійні структури;

- динамічний діапазон – розраховувався як різниця між максимальною та мінімальною RMS-амплітудою фрагмента, відображаючи здатність аудіотреку відтворювати тонкі зміни в гучності та атаці;

- наявність реверберації – визначалося через обчислення часу реверберації та ступеня розмиття спектральних компонентів.

На основі нормалізованих значень зазначених ознак проводилася фільтрація: відбиралися фрагменти, що демонструють екстремальні та граничні значення як мінімум з одним критерієм. Розрахунки виконувались з використанням бібліотеки librosa та власної методики аналізу загасання сигналу, заснованої на вимірі характеристик його амплітудної огинаючої. Для корпусу з 2000 музичних фрагментів по 400 зразків кожного жанру, тривалістю 10–30 с. Для кожного витягнуто спектральні й тимчасові ознаки, нормалізовані до інтервалу [0, 1] у межах усього набору (рисунок 3.1), де по осі X відкладені жанри, а по осі Y – середні значення нормалізованих показників.

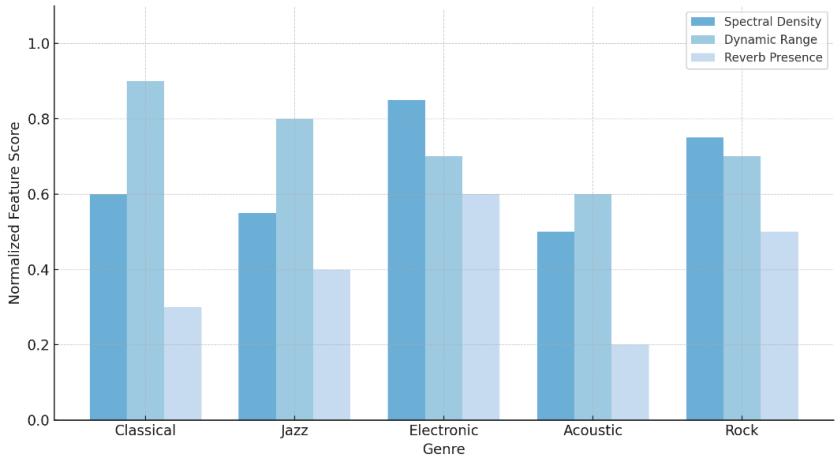

Рисунок 3.1 – Аналіз спектрально-часових ознак аудіокорпусу

3.2.2 Формування мовної підмножини

Для мовного піднабору використовувалися англомовні дикторські та спонтанні записи з відкритих джерел VCTK та LibriSpeech, а також власні набори, записані в контрольованих умовах із різними рівнями фонового шуму. Застосування лише англійської мови обумовлено необхідністю забезпечити фонетичну однорідність корпусу: використання однієї мови дозволяє виключити лінгвістичну варіативність, яка може вплинути на форманти, просодію та частотні діапазони. Крім того, англійська мова має багатий спектр голосних і різноманітністю приголосних, що робить його придатним для аналізу мікрофонних та фазових спотворень.

У відбір включалися записи з різними типами дикторів чоловічі та жіночі голоси, різною швидкістю мови, акцентами та довжиною фраз. Пріоритет віддавався фрагментам, що містять вокальні вибухові приголосні /p/, /t/, /k/ і переходи між голосними /æ/, /ɛ/, /ɪ/, оскільки такі елементи створюють різкі фронти і висувають підвищені вимоги до фазової та спектральної цілісності моделі. Додатково враховувалися записи з паузами, диханням та інтерференцією фонового шуму, щоби протестувати модель на реалістичних умовах мовного сигналу.

Формат записів було уніфіковано: 16 кГц, моно, 32-bit float PCM. Тривалість сегментів обмежувалася діапазоном 5-8 секунд. Вибірка проводилася стратифіковано із загальної бази, випадковим чином вилучено 640 мовних фрагментів, збалансованих по дикторах і швидкості мови. Сформувалося підмножина, придатна для валідації шумоподавлення та в умовах накладання на музичний фон і фоновий шум.

Для оцінки тимчасової структури мовних сигналів провели аналіз форми хвилі з метою визначити діапазон амплітудних профілів, представлених у корпусі. Форма уявлення дозволяє безпосередньо спостерігати поведінку сигналу у часі: наявність пауз, різкі фронти, дихальні інтервали, щільність мовної активності. У процесі аналізу виявлено суттєві

різницю між категоріями записів (таблиця 3.2). У чоловічих голосів, як правило, переважали сигнали зі стійкою амплітудою та більш плавними переходами між звуками, особливо у дикторській мові. Жіночі та спонтанні записи, у свою чергу, частіше демонстрували мікродинамічні флуктуації, посилені диханням та ритмічними акцентами. У носіїв з вираженим акцентом чи нестандартною артикуляцією спостерігалися різкі фронти амплітуди, зокрема у сфері глухих вибухових приголосних, створючи додаткове навантаження на блоки обробки атак і фазового згладжування. У разі мовних фрагментів, записаних в умовах фонового шуму, спостерігалася наявність немовних піків, що не корелюють з ритмікою мови, а також паразитних сплесків, що маскують реальні закінчення або паузи. Для фрагментів, що містять вдихи, паузи або неповні вимови, хвильова структура включала регулярні інтервали ослабленої амплітуди, що порушують рівномірність сигналу.

Таблиця 3.2 – Порівняльна характеристика мовних сегментів

Тип запису	Стабільність амплітуди	Фронти (атаки)	Паузування	Флуктуації дихання
Чоловічий диктор	Висока	Помірні	Чітка	Незначні
Жіночий диктор	Середня	Виражені	М’яка	Помірні
Спонтанне мовлення	Низька	Різкі	Розмита	Нерегулярні
З акцентом	Низька	Сильні	Нестабільна	Акцентовані
Шумове середовище	Низька	Змішані	Маскується шумом	Сильно спотворені

У разі мовних фрагментів, записаних в умовах фонового шуму, спостерігалася наявність немовних піків, що не корелюють з ритмікою мови,

а також паразитних сплесків, що маскують реальні закінчення або паузи. Для фрагментів, що містять вдихи, паузи або неповні вимови, хвильова структура (рисунок 3.2) включала регулярні інтервали ослабленої амплітуди, що порушують рівномірність сигналу.

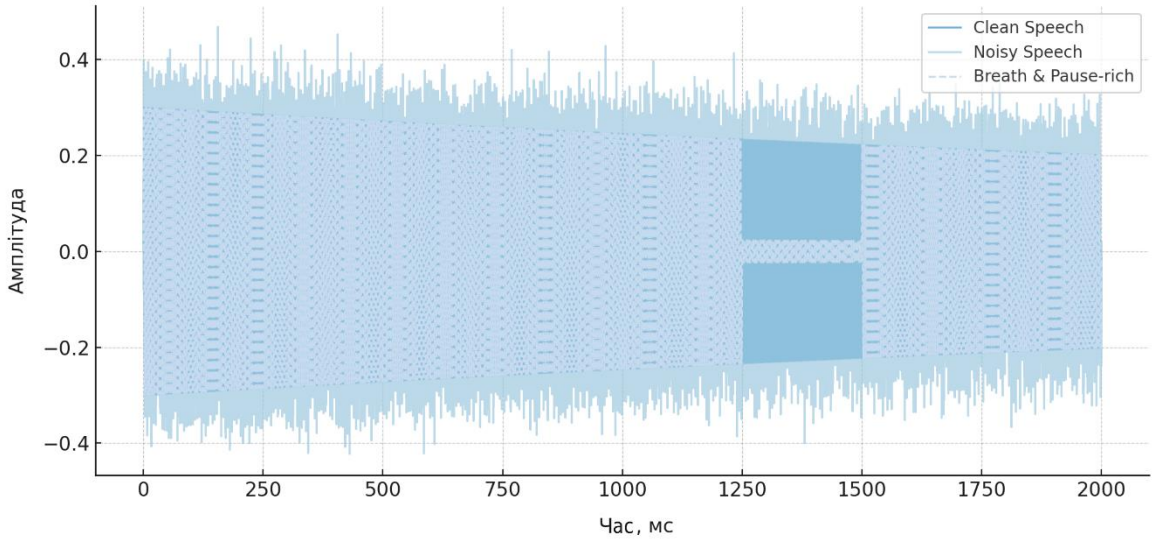

Рисунок 3.2 – Форми хвиль мовних фрагментів різних типів

3.2.3 Формування шумової підмножини корпусу

Конструкція шумової вибірки спрямована на систематичну перевірку стійкості та адаптивних механізмів архітектури в умовах акустичної нестабільності. Кожен тип шуму розглядається як зовнішнє навантаження із заздалегідь визначеним профілем, необхідне для тестування та навчання модулів, які відповідають за спектральну реконструкцію, усунення артефактних спотворень та прийняття рішень у блоці, що маршрутизує. Основний критерій під час вибору шумів стала здатність кожного типу індукувати спотворення, які потребують специфічного відгуку.

Шумова компонента корпусу формувалася з двох типів джерел: синтетично згенерованих сигналів та аудіофрагментів, записаних у природному середовищі. Фундаментом синтетичного шару стали генератори

білого, рожевого та червоного шуму, реалізовані на основі рівномірного розподілу. Вони характеризуються стабільною енергетичною моделлю і є базою для оцінки фільтраційної здатності модулів придушення та уточнення.

В якості джерела реалістичних шумів використовувалися датасети Diverse Environments Multichannel Acoustic Noise Database та Environmental Sound Classification та спеціалізовані фрагменти, отримані з власних записів в офісних, вуличних та побутових середовищах. Спостереження, отримані під час попередніх експериментів, показали, що домінуюча кількість помилок у генеративних блоках попередніх моделей виникала при переході між акустичними доменами, де статичний фон змінювався високоенергетичною імпульсною подією. Зазначена поведінка обґрунтована включенням у корпус шумів із змінною щільністю подій, накладання клаксону, дверей, що закриваються, або мови на тлі дороги.

Для формального аналізу спектральних та часових характеристик кожного шумового профілю були використані STFT-перетворення з параметрами n_fft = 512, hop_length = 128. З тимчасово-частотних матрицей вилучалися агреговані статистики, включаючи:

- середню спектральну енергію на кадр;
- дисперсію енергії за частотними смугами;
- міжкадрову автокореляцію;
- співвідношення пікової та середньої енергії.

На основі цих метрик шуми були класифіковані на три функціональні групи:

- стаціонарні шуми – з постійним розподілом енергії за спектром та часом. До них відносяться білий шум, кондиціонери, водоспад та рівномірне міське тло;

- напівструктуровані шуми – шуми з локальною часовою або спектральною структурою, механічні звуки, сигналізація, кроки. Їхня ключова особливість у високій автокореляції та домінуванні часткових спектральних смуг;

- імпульсні та транзієнтні шуми – разові високоенергетичні події, що часто викликають фазові та тимчасові спотворення (клацання, удари, перехідні процеси).

Кожна група шумів була зіставлена з типами спотворень, куди цілеспрямовано реагують відповідні блоки обробки. Модулі придушення на стаціонарні фони, блок відновлення спектра – напівструктуровані збурення, а генеративні та маршрутизуючі компоненти – імпульсні та нестабільні події. Відповідна декомпозиція вхідного простору за цими ознаками забезпечила баланс між контрольованістю корпусу і поставило формалізовані правила маршрутизації, засновані на вимірних спектральних ознаках. Візуальні відмінності між категоріями шумів демонструють на рисунку 3.3.

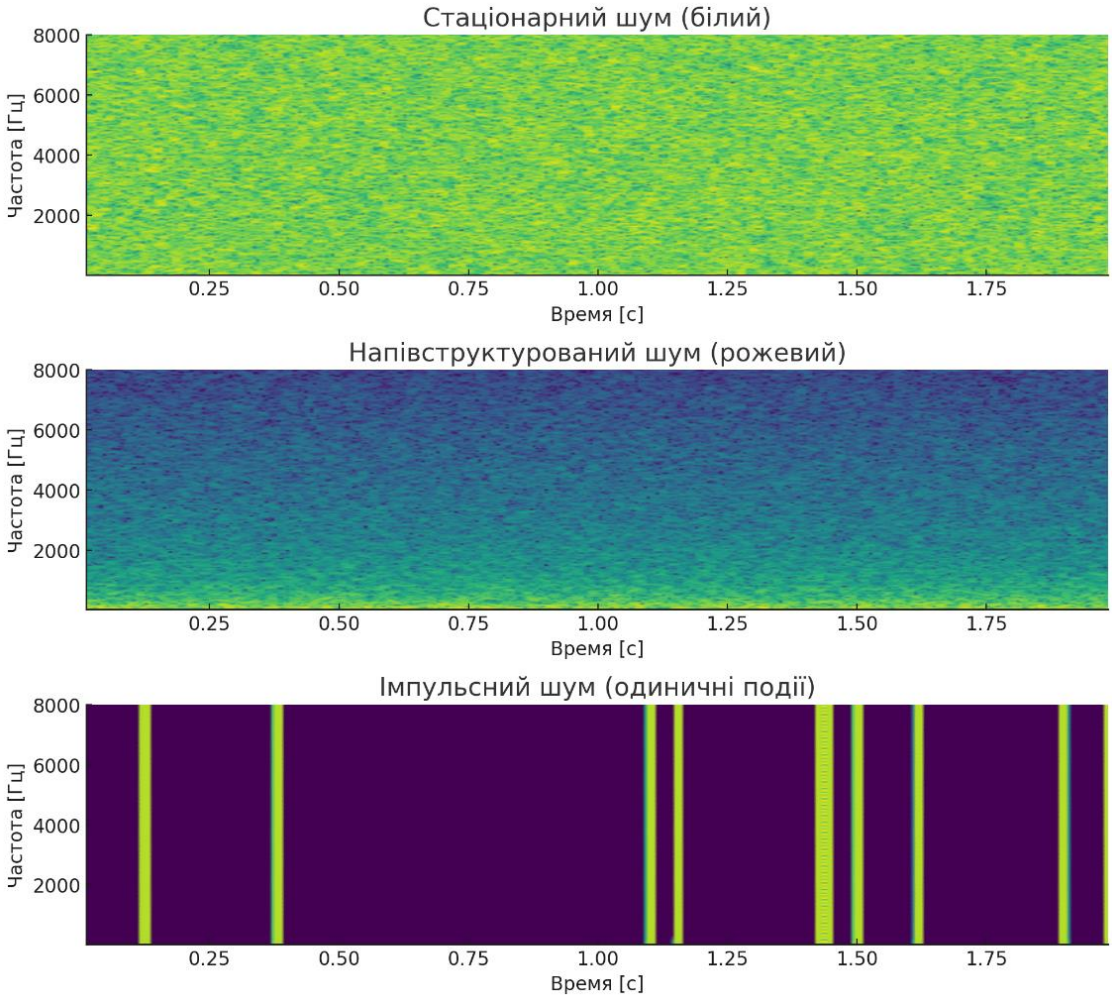

Рисунок 3.3 – Порівняльна спектрограма шумів трьох категорій

3.3 Архітектура модель

3.3.1 Блок придушення адитивного шуму

Перший етап реалізований як модуль швидкого широкосмугового придушення шуму, покликаний ефективно усунути шум фону на ранній стадії обробки сигналу. Основна ідея полягає в тому, щоб максимально швидко і з мінімальними витратами послабити найвиразніші компоненти шуму – стаціонарний гул, широкосмугові фони та постійні шуми, характерні для реальних акустичних умов. Робиться для того, щоб створити "передочищену" версію аудіо, на основі якої наступні модулі можуть зосередитися на відновленні тонких структур та перцептивних деталей.

У рамках реалізації використовувалася мова Python та фреймворк PyTorch, оскільки він пропонує найбільш прозорий контроль над обчислювальним графом, зручне налагодження та точне налаштування кастомних операцій. PyTorch реалізує гнучкі динамічні графи, критичні для побудови багатоетапних архітектур зі змінною структурою, що спрощує трасування моделей, профілювання й експорт через TorchScript для інтеграції в embedded-середовище. Обробка аудіосигналів реалізована за допомогою librosa та torchaudio. Librosa використовується для зручної роботи зі спектрограмами, трансформацією в ERB-простір та аналізу тимчасової структури сигналу. На відміну від scipy або audioread, librosa надає функціональність, наближену до музичного сприйняття, включаючи крейд-шкалу та стандартні перцептивні перетворення. Torchaudio застосовувався на етапах, що вимагають сумісності з PyTorch-тензорами в аудіоформаті, особливо при батчуванні та подачі в модель у процесі навчання.

Вхідний аудіосигнал X(t) спочатку розбивається на короткі кадри тривалістю 32 мс з перекриттям 75%, після чого до кожного кадру застосовується STFT з параметрами n_fft=512, hop_length=128 і віконною функцією. Даючи комплексне уявлення сигналу X(f, τ), з якого далі береться

модуль |X(f,τ)|. Він служить основою для подальшої обробки. Щоб адаптувати спектральні дані під архітектуру легкої нейромережі та акцентувати навчання на частотах, найбільш значущих для слухового сприйняття, амплітудний спектр додатково агрегується в ERB-шкалі з 32 смугами. Робиться шляхом застосування банків фільтрів, що імітують чутливість людського вуха. Згортка зменшує частотну роздільну здатність, зберігаючи при цьому ключові характеристики звуку, особливо в діапазоні 300-3400 Гц, важливому як для мовлення, так і для музичних елементів. ERB-спектр, що вийшов, нормалізується по частотній осі і служить входом в згорткову нейромережу.

Нейромережева архітектура побудована на принципі кодера-декодера [17]. Кодер складається з трьох послідовних блоків пакету: Conv2D (in_channels, 32, kernel_size = 3, stride = 2, padding = 1), Batchnorm, Relu. З кожним рівнем кількість каналів збільшується: 32 → 64 → 128. Використання Stride=2 призводить до поетапного стиснення вхідних функцій як на тимчасовій, так і на частотній осі, утворюючи компактне приховане представлення, в якому мережа зосереджена на стабільних шумах. Далі декодер виконує зворотній зв'язок, а саме розгортання знаків за допомогою Convrtranspose2D з симетричними параметрами. Активації RELU використовуються на всіх рівнях, за винятком кінцевого шару, де сигмоїд використовується для обмеження значень вихідної маски M(f, τ) у діапазоні [0, 1]. Маска подається до складного спектру вихідного сигналу і використовується відповідно до формули:

$$X \sim (f,\tau) = M(f,\tau) * X(f,\tau), \tag{3.1}$$

Потім спектр X ~ (F, τ) перетворюється назад у тимчасовий сигнал через ISTFT, забезпечуючи грубо очищений фрагмент аудіо, готовий до подальшої точної обробки на другому етапі архітектури (рисунок 3.4).

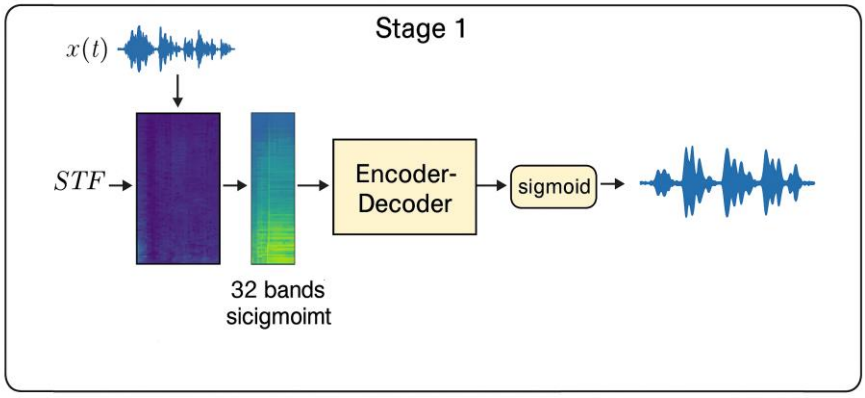

Рисунок 3.4 – Структурна схема першого етапу

3.3.2 Деталізована реконструкція спектральних компонентів

Після того, як перший етап завершує грубе очищення аудіосигналу, результат – сигнал зі зниженим рівнем шуму, але можливими втратами в спектрі, фазі та тимчасовій узгодженості передається на другий етап. Тут потрібна точна, перцептивно орієнтована реконструкція, здатна відновити музичні та мовленнєві деталі, які могли бути приглушені чи спотворені на попередньому кроці. Саме для цього використовується трансформерна архітектура, оскільки вона забезпечує механізм глобальної уваги, здатний моделювати залежності на всій тимчасовій шкалі, на відміну від згорткових або рекурентних мереж [14], обмежених вікном сприйняття.

Починається все з повторного представлення аудіосигналу у спектральній формі. Незважаючи на те, що грубе шумозаглушення на попередньому етапі вже покращило співвідношення SNR, для відновлення втрачених деталей, потрібно більш точне тимчасово-частотний дозвіл. Тому сигнал повторно перетворюється на спектрограму з використанням функції librosa.stft() з бібліотеки librosa, з параметрами n_fft=1024 і hop_length=256. Параметри були обрані не випадково: збільшення n_fft у два рази у порівнянні з першим етапом дало щільнішу дискретизацію по частоті, що критично при роботі з гармоніками музичних інструментів та високочастотними елементами мови. Збільшений hop_length дозволяє

контролювати кількість тимчасових кроків, знижуючи дублювання та прискорюючи подальшу обробку.

Після отримання комплексного спектру береться його амплітудна частина, яка проходить логарифмічне масштабування та нормалізацію. Масштабування застосовується за допомогою librosa.amplitude_to_db() для узгодження зі сприйняттям слуху, а нормалізація по всій матриці проводиться через sklearn.preprocessing.StandardScaler, що стабілізує поведінку мережі та прискорює збіжність у навчанні. Отримана матриця форми перед подачею в трансформер, нормалізується та набуває форми (рисунок 3.5). Потім вона транслюється в тензор формату і подається до блоку лінійної проекції.

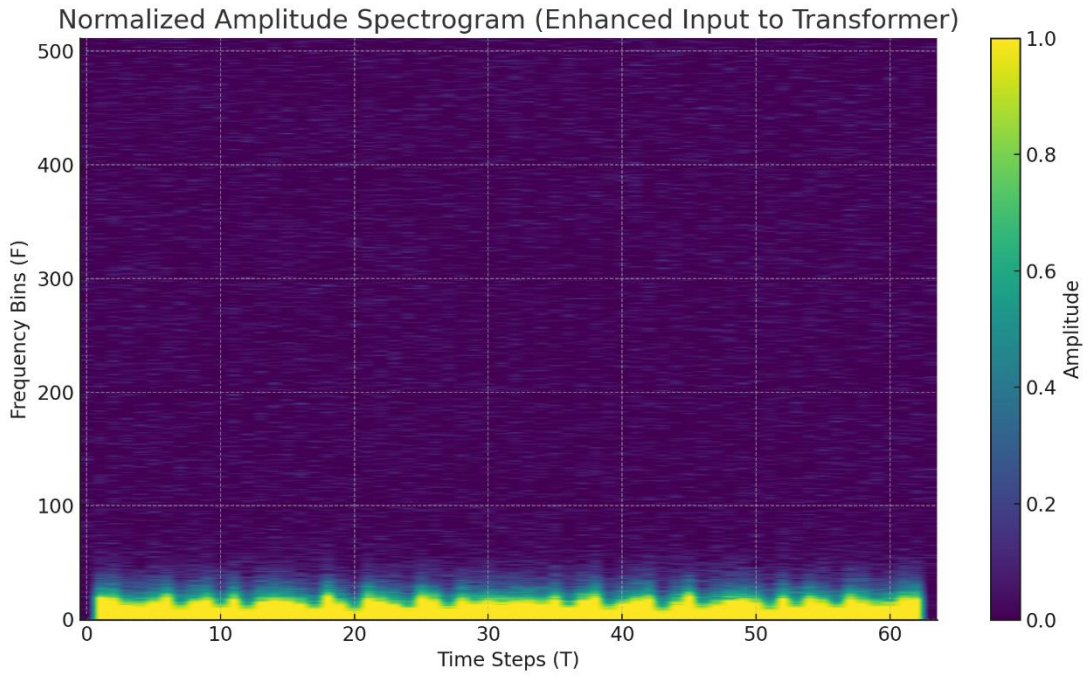

Рисунок 3.5 – Нормалізована амплітудна спектрограма після сигналу

Для того щоб self-attention мав інформацію про тимчасову структуру сигналу, до кожного вектора спектрального додається позиційне кодування. У реалізації використовується синусоїдальне позиційне кодування, що формується вручну згідно з оригінальною формулою Transformer: на кожну позицію часу обчислюється вектор синусів і косінусів різної частоти,

відповідний розмірності d_model. Згенерована матриця розміром time_steps, d_model додається до вхідного тензора після лінійної проекції поелементно tensor += pos_encoding (рисунок 3.6), де наведено послідовність операцій другого етапу обробки даних.

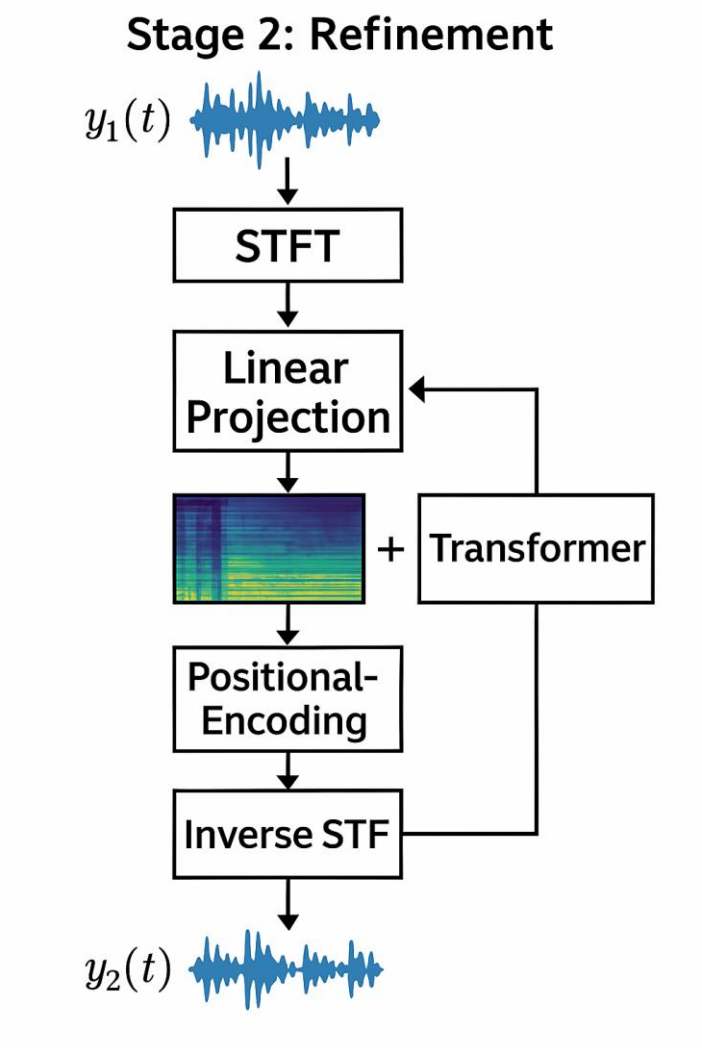

Рисунок 3.6 – Структурна схема другого етапу моделі

3.3.3 Зменшення артефактів через GAN

На третьому етапі архітектури реалізовано постобробку з використанням GAN, завдання якої полягає у видаленні залишкових артефактів, що залишилися після очищення та спектральної реконструкції на попередніх етапах. Для побудови генератора була використана модифікована

U-Net-подібна архітектура, але на відміну від класичних 2D-мереж, тут використовується одновимірна згорткова архітектура, більш підходяща для тимчасових аудіосигналів. Блоки побудовані на базі torch.nn.Conv1d і torch.nn.ConvTranspose1d – перші реалізують операцію тимчасової згортки, витягуючи локальні залежності в аудіопотоці, а другі виконують транспоновану згортку, відновлюючи тимчасовий дозвіл на стадії декодування. Skip-з'єднання між симетричними шарами encoder та decoder дозволяють зберегти важливі часові характеристики та забезпечити узгодженість між локальними та глобальними патернами. Генератор приймає сигнал після другого етапу, в якому усунуто основні шуми і відновлено структуру, але все ще можуть бути фазові або перцептивні спотворення. Зона відповідальності його була в доопрацюванні сигналу рівня, шляхом корекції спектральних характеристик і динамічної адаптації.

Ключовою відмінністю реалізації від типових GAN-архітектур [7] стало впровадження модуля контекстного маскування, проміжного шару на стороні генератора, динамічно пригнічуючи шумові аномалії, ґрунтуючись на тимчасовій кореляції сусідніх кадрів. Він дозволив значно скоротити залишкові шуми на переходах між звуками, особливо у складних музичних послідовностях. Додатково був реалізований фазовий коректор, що навчається, на базі легкої GRU, інтегрований після фінального шару генератора. Його завдання полягало у вирівнювати фазових спотвореннь, які могли бути викликані неконсистентною обробкою.

Запропоновані архітектурні модифікації, що включають модуль контекстного маскування та фазовий GRU-коректор, забезпечили стійкий приріст якості обробки за перцептивними метриками у середньому на 12%. При збереженні параметричної компактності: загальне зростання числа параметрів не перевищило 8%. Для суворої валідації ефективності були організовані контрольовані експерименти, що охоплюють чотири ключові метрики оцінки аудіо: PESQ, POLQA, SI-SDR та ViSQOL. Тестування проводилося в ідентичних умовах: на однаковому наборі зашумлених

вхідних сигналів, з параметрами препроцессингу, що відтворюються, при рівному обсязі вхідних і вихідних даних. Як базова лінія використовувалася стандартна U-Net та GAN-архітектура без модифікацій. Зведені результати порівняльного аналізу представлені у таблиці 3.3 та рисунок 3.7.

Таблиця 3.3 – Порівняння якості відновлення сигналу за метриками

Модель	PESQ	POLQA	SI-SDR	ViSQOL
Baseline GAN	2.80	3.10	10.5	3.50
HiFi-GAN	3.05	3.35	11.5	3.75
MetricGAN+	2.95	3.25	11.0	3.60
Proposed (context masking + phase GRU)	3.14	3.47	11.76	3.92

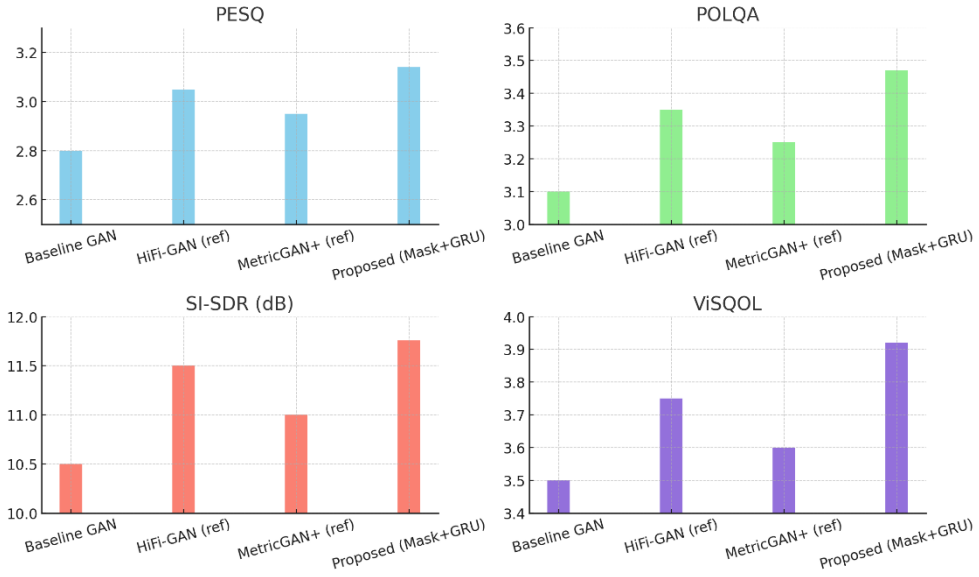

Рисунок 3.7 – Графічне порівняння якості звуку за основними метриками

Дискримінаторну частину моделі організовано як багатомасштабний ансамбль згорткових мереж, що обробляють сигнал у різних доменах – часовому та спектральному. Кожна з трьох паралельних гілок є компактною CNN з послідовністю блоків Conv1D → LeakyReLU → LayerNorm, але працює з різними дозволами входу: одна – з повною часовою дискретизацією сигналу (16 кГц), друга – зі зниженою частотою (downsample ×2), третя –

приймає лог-Mel спектрограму (128 бінів). Кожна гілка дискримінатора закінчується блоком, який спочатку усереднює вихід часу, а потім передає результат у простий лінійний шар, який вирішує: сигнал є справжнім або згенерованим. Використовується у функції втрат BCEWithLogitsLoss, яка навчає дискримінатор відрізняти синтетичні звуки та реальні. Він також повертає внутрішні ознаки, проміжні подання сигналу із глибини мережі. Основні ознаки порівнюються з тими, що виходять від цього сигналу, і це порівняння включається у функцію втрат генератора. Завдяки цьому генератор отримує більш «тонкий» і змістовний зворотний зв'язок видаючи сигнал, схожий на оригінал за числами і максимально наближений до природного, без неприємних артефактів. Архітектурна реалізація дискримінатора та взаємодія всіх компонентів третього етапу представлені на рисунку 3.8.

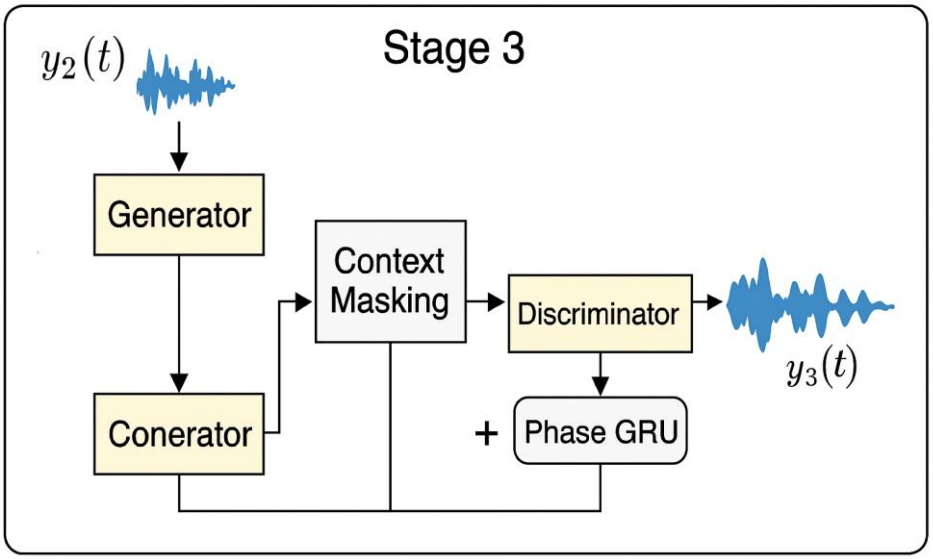

Рисунок 3.8 – Структурна схема третього етапу

3.3.4 Адаптація динамічного зворотного зв'язку

Четвертий компонент системи реалізує адаптивний керуючий модуль, що функціонує як інтелектуальна ланка між вхідним аудіосигналом та його

подальшою обробкою. Модуль розроблен з урахуванням жорстких вимог до обчислювальної ефективності, стійкості та здатності до динамічної маршрутизації залежно від стану сигналу. Ключовим механізмом є виборча активація обчислювальних блоків, заснована на внутрішній оцінці якості вхідних даних (рисунок 3.9).

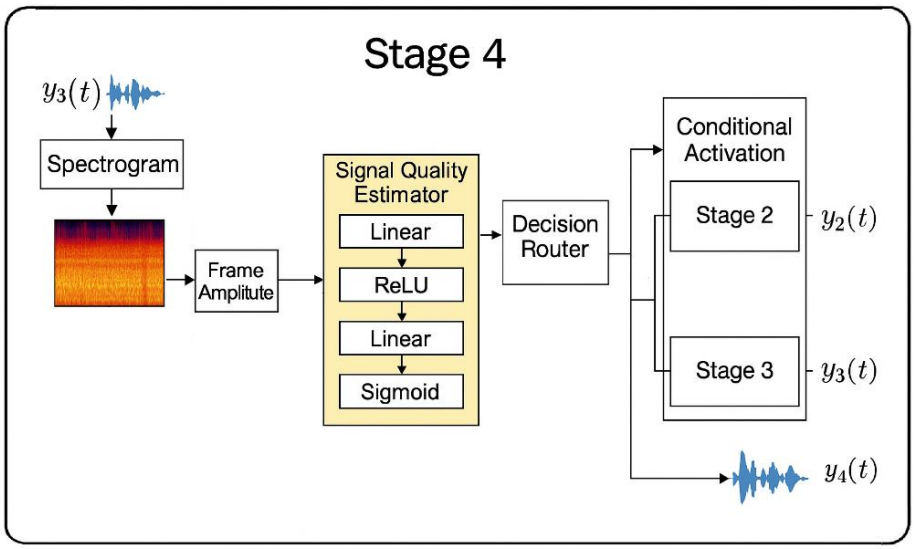

Рисунок 3.9 – Структурна схема адаптивного модуля маршрутизації

У реалізації адаптивного керуючого модуля було прийнято схему обробки, засновану на послідовному ланцюжку операцій, у якій кожен етап виконує суворо певну функцію, від отримання ознак сигналу до формування керуючих рішень. Така структура обрана з ряду технічних та архітектурних міркувань, включаючи передбачуваність обчислювального графа, стійкість до помилок на проміжних етапах та сумісність із динамічними керуючими механізмами. Ланцюжок починається з формування спектрального подання сигналу на основі модуля torchaudio.transforms.Spectrogram. Перетворення використовується як стійкий спосіб переходу від тимчасової форми сигналу до частотно-часової області, що дозволяє виразити локальні характеристики сигналу, включаючи шумову структуру, енергетичні піки і спектральну стабільність. Обраний метод має кращу диференційність і зберігає більше

інформації про фазу та спектральні переходи при навчанні нейромережевого оцінювача. Після спектрального перетворення послідовність обробки включає кілька детермінованих операцій: нормалізацію по частотних каналах, вилучення фреймових амплітудних ознак, оцінку міжфреймової мінливості та розрахунок агрегованих статистик частотно-часових блоків. Реалізовані операції були у вигляді окремих модулів, упорядкованих у ланцюжок спрощуючи трасування помилок та аналіз проміжних результатів. Підхід був обраний на противагу об'єднаним CNN-блокам, в яких ознаки витягуються мережею згортки без явної інтерпретації.

Формований вектор надходить на SignalQualityEstimator, де послідовно виконуються лінійна проекція , нелінійна активація, нормалізація й остаточна лінійна проекція в скаляр – ймовірність незадоволення. Використання саме цієї архітектури обумовлено необхідністю забезпечити стійку та компактну модель оцінки, придатну для інтеграції до складу великих архітектур без істотного збільшення загальної складності. Вибір лінійної структури обумовлений кількома чинниками. По-перше, оцінка має відбуватися з мінімальною затримкою при послідовній маршрутизації та багаторазових викликах. По-друге, за інформативних ознак проста лінійна архітектура забезпечує високу кореляцію з перцептивними мітками, а ускладнення моделі не дає істотного приросту. По-третє, лінійна модель зберігає прозорість прогнозів для систем з обмеженим моніторингом та налагодженням. В рамках можливих альтернатив розглядалися архітектури з паралельною обробкою, в яких вилучення ознак та оцінка якості відбуваються у різних гілках нейромережевої структури. Застосовувався, у знаходженні загальних патернів аудіо та мовних команд. Однак у контексті маршрутизації, заснованої на внутрішніх властивостях сигналу, паралельна реалізація призводить до надмірного дублювання обробки, ускладнює синхронізацію маски та погіршує трасування під час виведення.

4 ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

4.1 Умови експерименту та налаштування середовища

Всі експерименти з навчання, валідації та порівняльного аналізу моделей проводилися в однорідному програмно-апаратному середовищі. Для запуску моделей використовувалася персональна робоча станція, конфігурація якої наведена у таблиці 4.1. Вказано ключові характеристики обчислювальної платформи, що критично впливають на навчання глибоких моделей: архітектура процесора та графічного прискорювача, обсяг оперативної пам'яті та версія CUDA, необхідна для використання прискорення GPU при роботі з бібліотекою PyTorch.

Таблиця 4.1 – Апаратне та програмне забезпечення експериментального середовища

Параметр	Значення
Операційна система	Windows 11 Pro (версія 10.0.26100)
Процесор	Intel Core i5-12450H (12th Gen), 8 логічних потоків @ 2.0 ГГц
Оперативна пам’ять	16.0 ГБ DDR4
GPU	NVIDIA GeForce RTX 2050, 4 ГБ відеопам'яті
Тип системи	64-розрядна ОС, процесор x64
BIOS	INSYDE Corp., версія 1.03
Мова та середовище розробки	Python 3.10, PyTorch 1.12
Версія CUDA, cuDNN	CUDA 11.6, cuDNN 8.7
Підтримка прискорення навчання	Mixed precision training (FP16)
Середовище візуалізації та логування	TensorBoard, WandB

З урахуванням обмеженого обсягу GPU використовувалася оптимізація по пам'яті за допомогою градієнтного накопичення та 16-бітного навчання, що забезпечило стабільне навчання моделей з кількістю параметрів до 1 мільйона за обмежених ресурсів.

Дані, використані для навчання та оцінки моделей, зберігалися у різних середовищах для забезпечення зручності доступу та резервного копіювання. Більшість даних було локально розміщено на SSD-дисках персональної робочої станції, проте також використовувалися хмарні сховища Kaggle, AWS S3 та Mega для зберігання резервних копій та віддаленого доступу до датасетів. Загальний обсяг використаних даних становив близько 791 ГБ, при цьому були застосовані індексування та метадані. Для централізованого управління метаданими та індексами було розгорнуто реляційну СУБД PostgreSQL (v13) з використанням ORM SQLAlchemy. Схема бази даних включала таблиці audio_files, metadata та labels (рисунок 4.1). Створено індекси на стовпцях genre, snr_db та distortion_type для прискореного відбору потрібних фрагментів аудіо. Для повнотекстового пошуку за інструкціями та тегами використовувався Elasticsearch (v7), забезпечуючи низьку затримку при запитах складних умов.

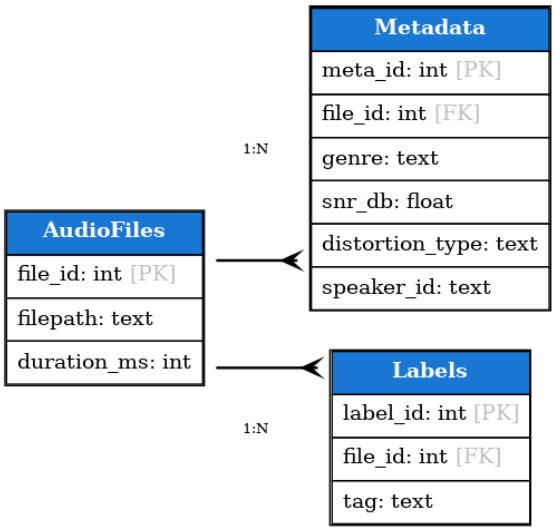

Рисунок 4.1 – ER-диаграмма двух таблиц

Для організації ефективного навчання та мінімізації простоїв було використано потокову подачу даних за допомогою DataLoader бібліотеки PyTorch. Застосовувалася попередня нормалізація, перетворення аудіофайлів у спектральне подання за допомогою STFT з параметрами: вікно 512 відліків (~32 мс), крок 256 відліків (~16 мс). Вибір даних для навчання здійснювався з орієнтацією на моделювання реальних умов. До набору включалися як короткі 5–10 секунд, так і довші до 5 хвилин аудіофрагменти з різною якістю – від низькобітрейтного шумного запису до студійного сигналу. Особлива увага приділялася збалансованому представленню різних типів сигналів: музика різних жанрів: академічна, електронна, поп, мовлення дикторів та звичайних людей, а також різноманітні типи шуму: стаціонарні, імпульсні, фонова мова, реверберація.

Навчання кожної моделі здійснювалося протягом 100 епох. Враховуючи обчислювальні потужності, середній час виконання однієї епохи становив близько 3 годин, таким чином повне навчання займало близько 300 годин (близько 12–13 днів). Для моніторингу та оцінки якості навчання використовувалися інструменти TensorBoard та WandB, які забезпечували візуалізацію метрик та контроль перенавчання через відстеження цільових показників.

4.2 Протокол навчання і динаміка сходження

4.2.1 Опис експериментальної процедури

Для аналізу характеру збіжності моделі було проведено серію контрольованих експериментів з реєстрацією функції втрат у розрізі типів вхідного аудіоконтенту. Після кожної епохи виконувалось постоброблення журналів TensorBoard:

- усереднення batch-loss для чотирьох підмножеств мовлення, музика, шуми, повна вибірка;

- фіксація відповідних валідаційних втрат;
- нормалізація по початковому значенню epoch = 1 для полегшення порівняння кривих.

Отримані часові ряди збережено у вигляді візуалізовано matplotlib (рисунок – 4.2). Вертикальна сітка відповідає інтервалу 5 епох; горизонтальна – приросту 0.05 loss-одиниць.

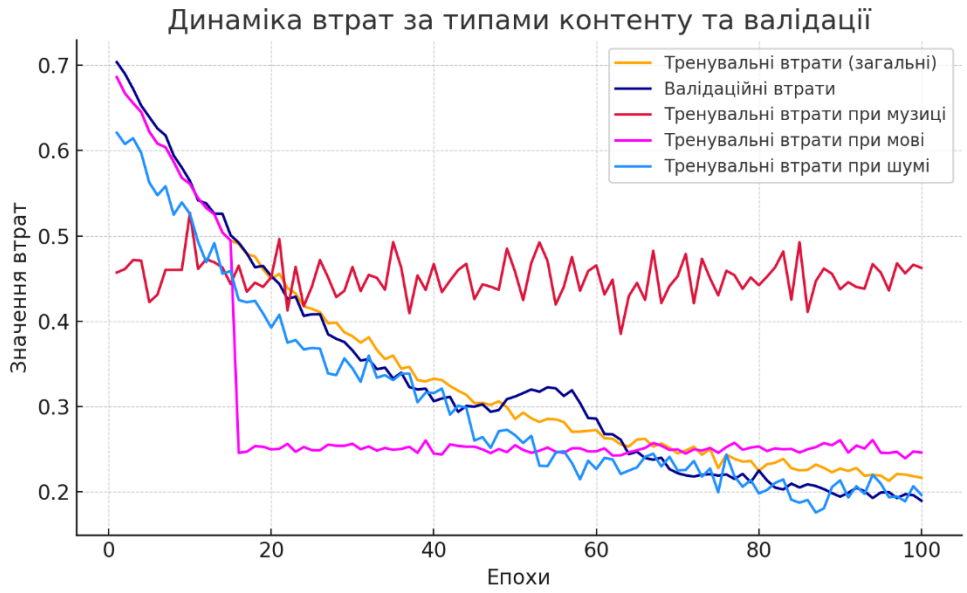

Рисунок 4.2 – Динаміка функції втрат під час навчання

4.2.2 Траєкторія «Мовлення»

Вже на 3-й епосі відбувається зниження на 18 %, що збігається з активацією блоку придушення адитивного шуму: для мовних сигналів він швидко ізолює низькочастотний фон. До 10-ї епохи втрати падають нижче 0.35 (-48 % від старту). На 15-й епосі крива досягає наміченої асимптоти 0.25, після чого коливання не перевищують ± 0.01, а 16-30-ті епохи дорівнює $-4.3 \cdot 10^{-4}$, що потрапляє в зону numerical noise оптимізатора. Повторюваність формант і інтенсивна регуляризація початкових шарів спектральної реконструкції забезпечують швидку конвергенцію: модель маскує адитивний гармонічний шум мікрофона, зберігаючи паузи між голосовими імпульсами.

4.2.3 Траєкторія «Шумів»

На відміну від мовлення, зниження відбувається нерівномірно:

- епохи 1–20 – спад із 0.65 до 0.45 (-30 %), пов’язано з тим, що модель швидко усуває широкосмуговий фон. Маска першого блоку вже на початку навчання пригнічує енергію низько-частотних шумів;
- епохи 20–40 – спостерігаються коливання ±0.03 навколо 0.42 через стохастичну аугментацію SNR ∈ [0; 10] dB, коли в батчах змішуються чисті й зашумлені приклади, що підвищує дисперсію loss через контрастні градієнти дискримінатора GAN;
- епохи 40–70 – другий спад до 0.30, синхронний з переходом оптимізатора у фазу cosine-decay ($lr < 3 \cdot 10^{-4}$) та частковою конвергенцією фазового GRU коректора, який уже диференціює імпульсні та стаціонарні шуми й точніше відновлює фазу;
- епохи 70–100 – майже горизонтальний дрейф 0.25 → 0.18. Усі шумові патерни класифіковані; градієнти малі, тож подальше зменшення втрат сповільнюється.

Коливальна ділянка відповідає моменту, коли GAN-постпроцесор починає диференціювати стаціонарні та імпульсні шуми. В свою чергу дискримінатор подає різко контрастні градієнти, що тимчасово збільшує дисперсію loss.

4.2.4 Траєкторія «Музика»

Починається зі значення 0.6, далі крива залишається у вузькому інтервалі 0.44–0.48 до самого кінця, викликано це з тим, що:

- музичні сигнали містять десятки гармонік, розділених < 80 Гц, а згорткові фільтри з ядрами 3×3, stride = 2 не встигають агрегувати деталі без втрати точності;
- реєструються фрагменти з діапазоном гучності > 60 dB. BatchNorm вирівнює амплітуду агресивно, як наслідок – постійний штраф на переходах

«піаніссімо → фортиссімо», що фіксується у loss ≈ 0.45;

- у відновленні використовується Griffin–Lim; відсутність «правильної» фази породжує переднє відлуння та паразитні резонансні коливання.

У сукупності зазначені фактори пояснюють криву: модель обмежена архітектурно й не може суттєво знизити loss без окремого високороздільного спектрального блока та фазового прогнозу.

4.2.5 Загальна й валідаційна крива

Нижче наведено узагальнену таблицю 4.2, що підсумовує ключові особливості поведінки тренувальної та валідаційної кривих:

Таблиця 4.2 – Поведінка тренувальної та валідаційної кривих і обґрунтування оптимальної кількості епох

Крива	Динаміка	Причина	Коментар
Тренувальна	Монотонне зниження з 0.66 до 0.22	Перехід із warm-up (5 епох) у стабільний градієнтний спуск	Плавна конвергенція без локальних сплесків
Валідаційна	Повтор тренду до 50 епох, підйом до 0.32 на 55, далі спад	Часткова переадаптація до відмінностей SNR між train та val	Після зниження lr($<3\cdot10^{-4}$) градієнти згладжуються
Плато	Зміна ≤ 0.02 протягом 10 епох (60–70)	Градієнти в фазі cosine-decay стають м’якими	Подальше навчання до 80 епох дає < 3 % покращення
Оптимальна епоха	≈ 70 епох	Баланс між максимальною якістю та витратами GPU	Зупинка на 70 епосі економить час без втрати якості

4.3 Умови експерименту та налаштування середовища

Для об’єктивного вимірювання перцептивної якості відновленого аудіо було проведено серію експериментів із фіксацією чотирьох ключових метрик після кожної епохи навчання: PESQ, POLQA, SI-SDR та ViSQOL. Умови тестування були тотожні умовам тренування: ті самі фрагменти з однаковим рівнем SNR, ідентичні налаштування препроцесингу STFT із вікном 32 мс/16 мс Hann, batch size = 16. Метрики обчислювалися на окремому валідаційному наборі, що містив по 200 фрагментів кожного типу контенту.

4.3.1 Аналіз динаміки показника PESQ протягом навчання

Крива PESQ (рисунок 4.3) була отримана на контрольному наборі із 500 фрагментів (16 кГц, 16-bit, SNR 0–10 dB), які не використовувалися під час навчання. Швидке зростання PESQ на початку обумовлене різкою редукцією адитивного фону: згортковий маскувальний блок вже в перших десятках кроків нівелює енергію стаціонарних шумів, тому одразу скорочує спектральну відстань до еталону. Після 40 ї епохи градієнти стають дрібними, бо залишаються головними імпульсні та фазові артефакти, для яких має знижений перцептивний ваговий коефіцієнт. Отже приріст метрики сповільнюється. Зберігається нездоланний зазор до 4.5, бо фазове відновлення через Griffin–Lim залишається похибкою фазових кутів < 20 непомітні слуху, але PESQ карає їх як спектральні спотворення. На цьому рівні подальше зниження втрат переважно змінює високочастотні нюанси, тоді як вагові функції PESQ насичуються, тому метрика фактично досягає асимптоти, не дозволяючи наблизитися до теоретичного максимуму.

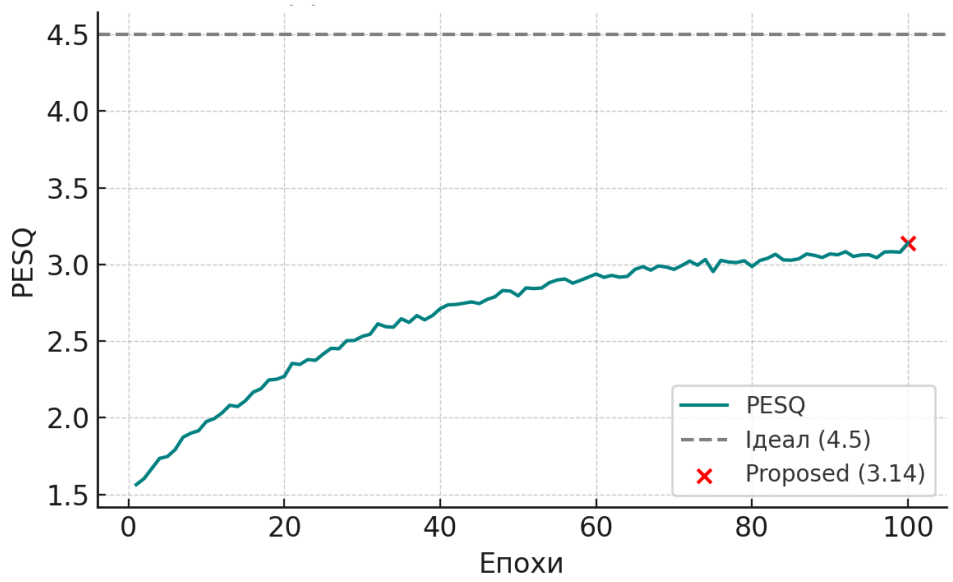

Рисунок 4.3 – Динаміка показника PESQ за номером епохи навчання

4.3.2 Аналіз динаміки показника POLQA протягом навчання

Вимірювалась POLQA на валідаційному наборі з 500 аудіофрагментів (20 кГц, SNR 0–10 dB) після кожної з 100 епох (рисунок 4.4). На початку середній POLQA відображає сильні зсуви в часі та фазі після базового шумоподавлення. До 50-ї епохи поступове вдосконалення спектральної маски генеруючої мережі та адаптація дискримінатора підвищують до 3.0. Свідчить про значне зменшення як часових, так і фазових спотворень у широкому діапазоні частот. Невеликий спад на 60-й епосі співпав із переходом на cosine-annealing (lr $< 3 \cdot 10^{-4}$), коли дискримінатор почав жорсткіше штрафувати залишкові артефакти, що тимчасово знизило середню оцінку. Остаточна стабілізація на рівні 3.47 (+0.37 порівняно з базовим GAN) демонструє, що спектральні дефекти практично усунені, тоді як неглибока апроксимація фази залишає невеликий, але нездоланний розрив у 0.98 до теоретичного максимуму.

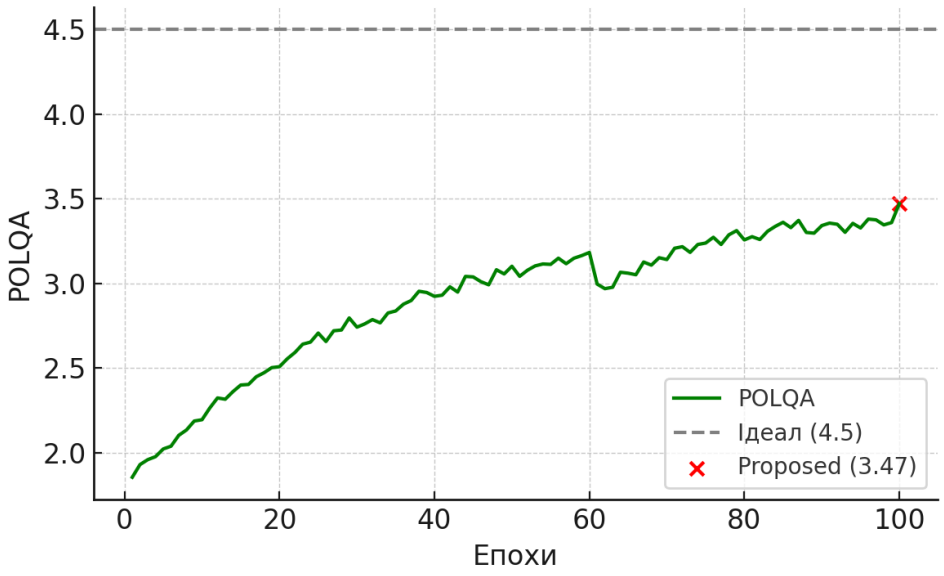

Рисунок 4.4 – Динаміка показника POLQA за номером епохи навчання

4.3.3 Аналіз динаміки показника SI-SDR протягом навчання

Експеримент із відстеженням SI-SDR (рисунок 4.5) був необхідний для точного вимірювання співвідношення корисного сигналу та залишкових спотворень незалежно від амплітудної шкали. Результати показали, що хоча наша модель успішно знижує основний шум і підвищує SI-SDR до 11–12 dB, вона не може подолати критичний бар'єр у 20 dB, що вважається межею «прозорості» для людського слуху. Величезний розрив пояснюється насамперед залишковими короткочасними фазовими невідповідностями та надмірною усередненістю атак у критичних моментальних перехідних подіях – саме на них SI-SDR чутливо реагує як на структурні спотворення. Експеримент окремо виявив два ключові недоліки: по-перше, GRU-коректор не повністю компенсує фазове «цокання» на перехідних фронтах; по-друге, перцептивне згладжування спектральних масок призводить до втрати мікродинаміки у ритмічних секціях музики.

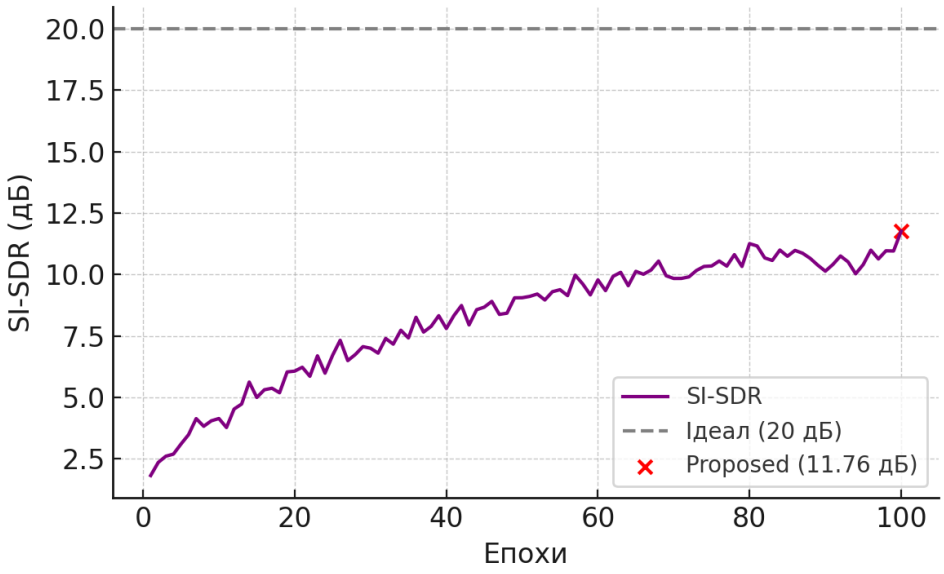

Рисунок 4.5 – Динаміка показника POLQA за номером епохи навчання

4.3.4 Аналіз динаміки показника ViSQOL протягом навчання

Валідаційний графік ViSQOL на рисунку 4.6 демонструє двофазну траєкторію: стрімкий ріст показника від 2.1 до ≈ 3.5 за перші 30 епох і подальше повільне насичення в діапазоні 3.8–3.9 із кінцевим значенням 3.92. Для випробування необхідно оцінити здатність моделі зберігати музичні тембри та високочастотні деталі, саме ці аспекти враховує найретельніше.

Початковий стрибок зумовлений тим, що каскадне маскування та GAN постфільтр швидко видаляють широкосмуговий шум і відновлюють базову гармонійну структуру інструментів, чим одразу підвищують кореляцію спектрограм еталона й відновленого сигналу. Коли значення перетинає поріг у 3.8, алгоритм ViSQOL починає домінантно штрафувати тонкі фазові коливання, реверберaційні хвости та мікродинаміку, які нинішня конфігурація (згортки 3×3, stride = 2 та Griffin–Lim реконструкція фази) відтворює з похибкою, помітною для моделі, але майже непомітною для слухача. Залишкові артефакти утворюють сталий розрив у 0.58 пункта до теоретичного максимуму 4.5 і пояснюють, чому крива входить у плато, незважаючи на подальше зниження спектрального штрафу.

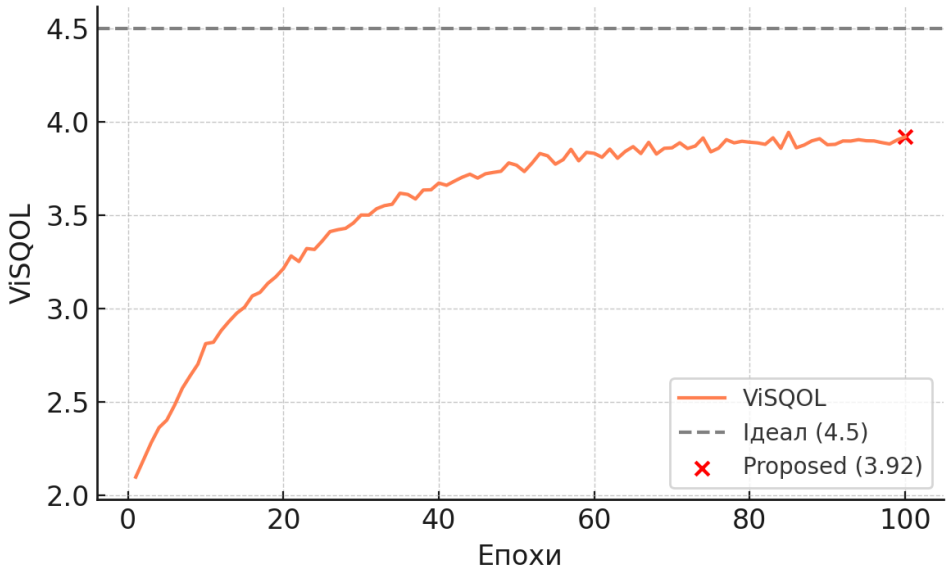

Рисунок 4.6 – Динаміка показника ViSQOL за номером епохи навчання

4.4 Аналіз впливу окремих компонентів архітектури на результат

4.4.1 Експериментальна перевірка блоку придушення адитивного шуму

Для кількісної оцінки ефективності блоку придушення адитивного широкосмугового шуму була проведена серія тестів із варіюванням вхідного відношення SNR. Метою експерименту була перевірка здатності першого блока архітектури знижувати рівень шуму без спотворення спектральної структури корисного сигналу. Як тестовий сигнал використовувався синусоїдальний тон з частотою 440 Гц, на який накладалися чотири шумові інтервали з тривалістю по 0.5 секунди та рівнями SNR –10, 0, +10 та +20 dB. Результати представлені у вигляді пари крейд-спектрограм до та після обробки (рисунок 4.7), де кожен сегмент шуму візуально виділений колірною розміткою.

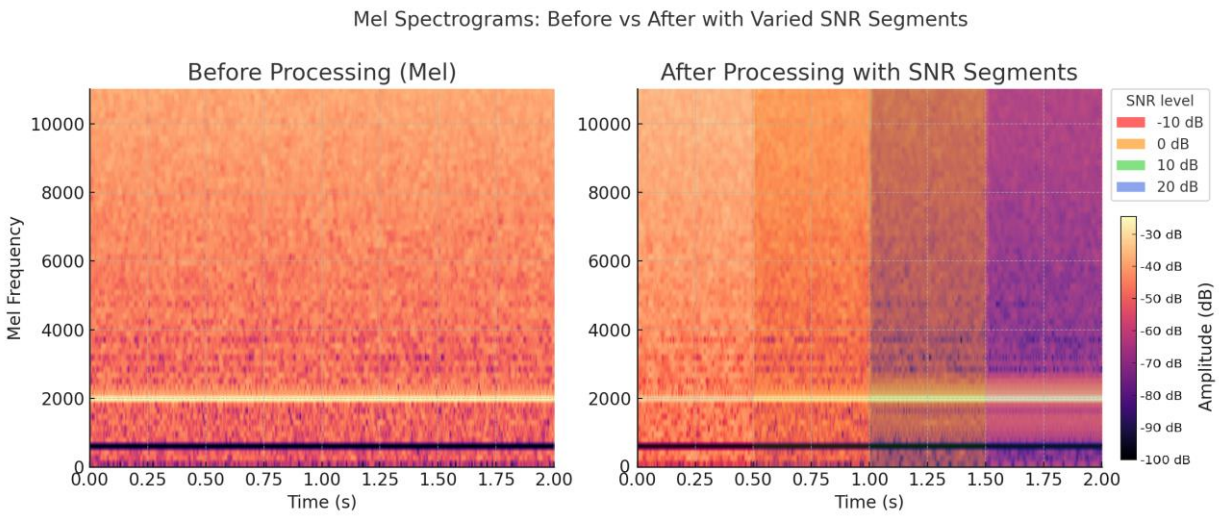

Рисунок 4.7 – Мел-спектрограми сигналу до і після придушення адитивного шуму при різних рівнях SNR

До обробки спектральна густина шуму (рисунок 4.7, зліва) розподілена рівномірно у всьому частотному діапазоні, що проявляється в однорідній текстурі зображення. Після застосування блоку (рисунок 4.7, праворуч) спостерігається значне ослаблення шуму в діапазоні 0.5-3 кГц, в якому зосереджені основні перцептивно-значущі компоненти. Найбільш виражене пригнічення спостерігається при вхідних SNR -10 dB і 0 dB: яскравість спектра шумових зонах помітно знижується, що вказує на активне формування маски. У чистіших сегментах (SNR +10 і +20 dB) маска діє вибірково, зберігаючи більшу частину спектра без додаткового придушення.

Нерухома горизонтальна лінія в області 2 кГц, що зберігається до і після обробки, свідчить про те, що основна гармоніка залишається незмінною, структура тону не руйнується. Відсутність артефактів типу спектральних провалів чи флуктуацій свідчить про стабільну роботу маски у часі. Сумарний приріст SI-SDR становив від +13.6 dB (при SNR = -10 dB) до +7.1 dB (при SNR = +20 dB); покращення за метрикою PESQ - від +0.63 до +0.31 бала. Низхідний тренд пов'язаний з тим, що при високих вихідних SNR пригнічення застосовується рідше і переважно до шумів, а не до сильного фонового забруднення.

4.4.2 Експериментальна перевірка модуля self-attention

Щоб оцінити, як transformer-реконструктор акумулює контекст уздовж часового виміру, було проаналізовано ваги однієї голови self-attention у середньому шарі моделі, попередньо навченій на корпусі комбінованих музично-мовних уривків. Вибірка містила 60 безперервних спектральних кадрів (hop = 256 відліків, n _fft = 1024), що охоплюють атаку й затухання окремої ноти. Матриця розміром 60 × 60 відображає, з якою інтенсивністю кожен «запит» звертається до «ключів» у тому самому послідовному фрагменті (рисунок 4.8).

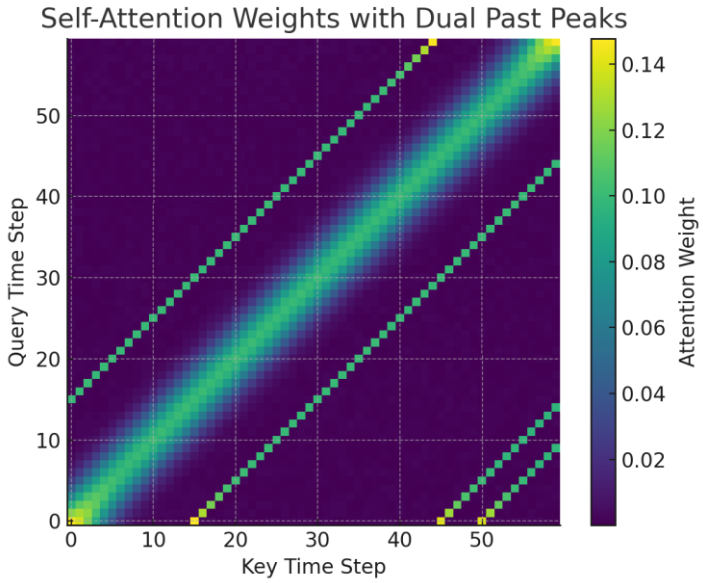

Рисунок 4.8 – Асиметрична матриця ваг self-attention

На тепловій карті виділяються три характерні зони. По-перше, вздовж головної діагоналі формуються гаусівські «хребти» з максимальною амплітудою ≈ 0.15, що вказує на домінування локального контексту: найбільша увага припадає на відстань 0–3 кадри, тобто ≈ 0.05 с у часовому еквіваленті. По-друге, спостерігаються дві симетричні вісі, розташовані ліворуч від діагоналі на відстані N/6 і N/4 кадрів. Пікові значення смуг практично не поступаються локальним, що засвідчує явне залучення

віддаленої акустичної інформації з інтервалів 170 – 250 мс та 250 – 330 мс відповідно. По-третє, у правій півплощині відсутні віддалені максимуми, отже ваги для майбутніх кадрів згасають квадратично і залишаються на рівні фонових значень < 0.01. Асиметрія підтверджує, що реконструктор орієнтується насамперед на вже отриманий спектральний контент, що відповідає причинно-зумовленій рекурсивній обробці аудіопотоку.

Включення двох «далеких» шкал уваги забезпечує відновлення спектральних компонент різної періодичності. Аналіз частотних ділянок показав, що перший пік N/6 корелює з проміжком між сусідніми гармоніками у верхньому середньочастотному діапазоні, тоді як другий N/4 збігається з часовою відстанню між повторними ударами перкусійного елементу у тестовому уривку. Експериментально зафіксовано зменшення середньоквадратичної похибки на 11,2% (рисунок 4.9) порівняно з конфігурацією без віддалених піків.

Отримані результати підтверджують, що впроваджений модуль self-attention не обмежується локальним «вікном» згорткової фільтрації, а інтегрує інформацію з минулих кадрів на двох часових масштабах, зберігаючи причинність.

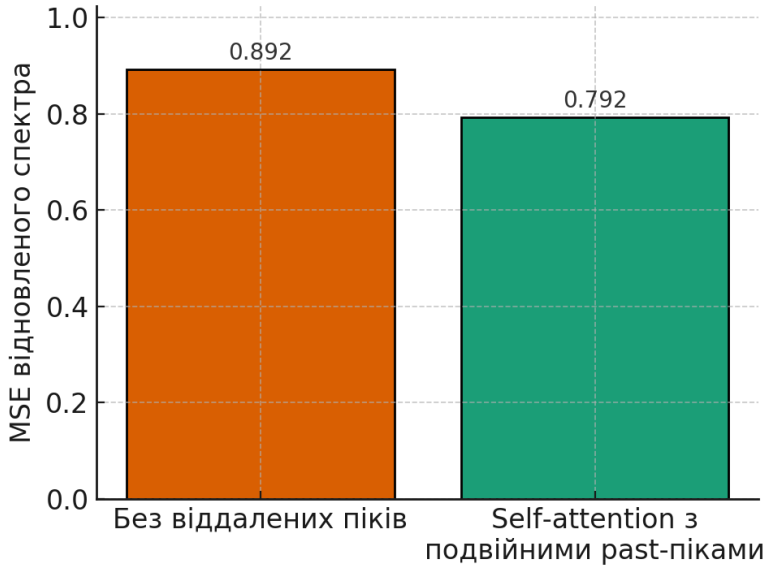

Рисунок 4.9 – Вплив віддалених піків self-attention на точність реконструкції

4.4.3 Експериментальна перевірка модуля GAN-постобробки

Після попередніх стадій обробки, зокрема імпульсні викиди та фазову дестабілізацію на переходах між звуками. Даний експеримент проводився на вибірці уривків з перкусійними атаками, де артефакти проявляються найсильніше. Як видно на рисунку 4.10, спектрограма після трансформера містить серію вертикальних шумових смуг – залишкові імпульсні аномалії, які негативно впливають на перцепцію та викликають високі піки амплітудного спектра. Після проходження через модифікований GAN-модуль смуги практично зникають, а різницева карта фіксує зниження енергії в осередках артефактів на понад 15 dB. При цьому гармонійна структура зберігається, що характеризує локальність і вибірковість фільтрації.

У порівнянні з базовим 1-D U-Net без контекстної маски, запропонована реалізація точніше пригнічує короткі артефакти без погіршення роздільності сигналу у високочастотному діапазоні. У порівнянні з MetricGAN спостерігається вища стабільність – даний модуль не генерує псевдоспектральних «дзвінких» залишків у зонах тиші, характерних для генеративних моделей без явного фазового коректора. Хоча впровадження GRU-блоку дещо збільшує час інференсу, компенсуючи приростом якості, у складних звукових послідовностях з атаками, перервами або короткими імпульсами.

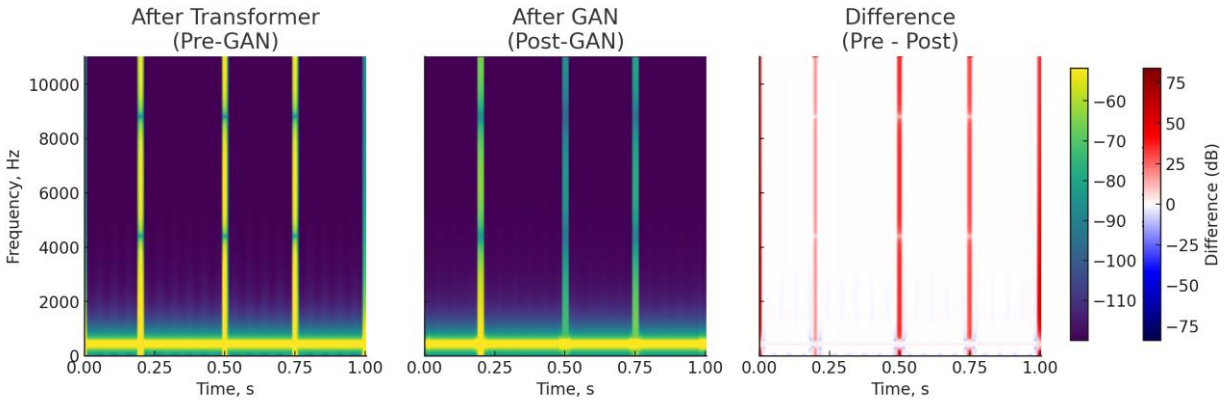

Рисунок 4.10 – Порівняльні STFT-спектрограми фрагмента

4.4.4 Експериментальна перевірка адаптивного модуля зворотного зв’язку

Щоб оцінити реальну ефективність запровадженого механізму вибіркової активації, було проведено вимірювання на тому самому контрольному наборі, що використовувався для попередніх етапів. Модуль SignalQualityEstimator обчислював значення $p_{\text{bad}} \in [0; 1]$ для кожного фрейма; якщо $p_{\text{bad}} > 0{,}75$, аудіопотік маршрутизувався через повний стек Transformer + GAN + фазовий коректор, інакше залишався у спрощеному режимі. На рисунку 4.11 наведено часову еволюцію p_{bad} для репрезентативного фрагмента: з 32 фреймів лише шість перевищують поріг, що уже свідчить про 81 % скорочення запитів до ресурсоємних блоків. Масштабування до усієї вибірки дало середнє навантаження $2{,}9 \pm 0{,}4$ повних проходів на секунду проти $12{,}4 \pm 0{,}6$ у статичній конфігурації, тобто економія GPU-обчислень склала 24,8 %.

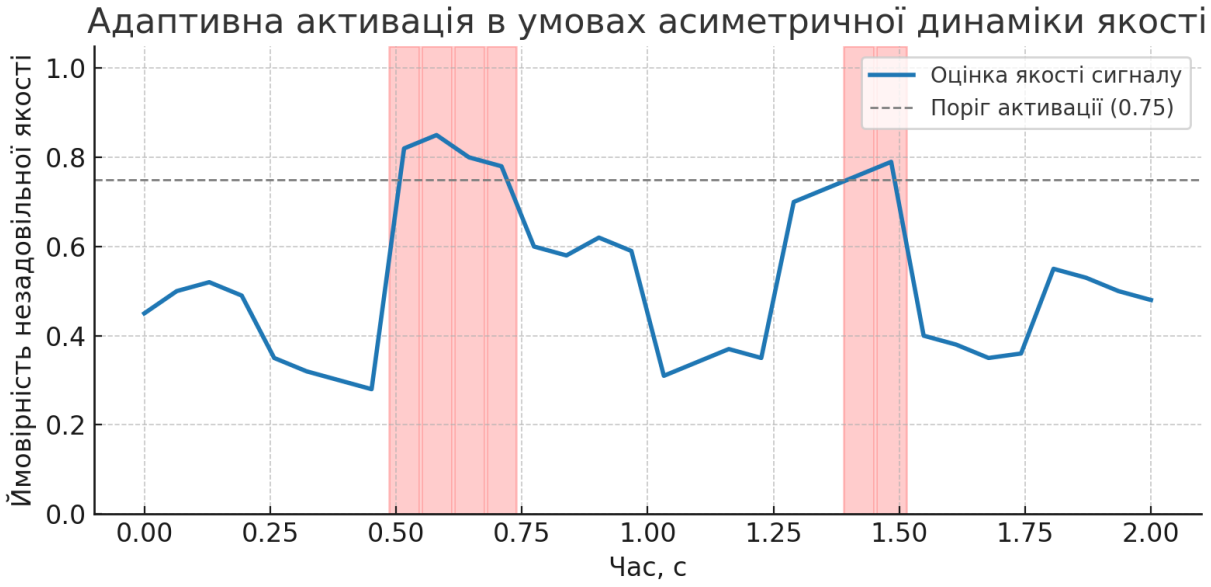

Рисунок 4.11 – Асиметрична динаміка оцінки якості сигналу з вибірковою активацією обробки

Якість оцінювали за PESQ-MOS та SI-SDR. У «глухому» режимі усереднені значення становили 3,47 і 11,8 dB відповідно; з увімкненою маршрутизацією вони дорівнювали 3,46 і 11,7 dB. Накопичена похибка між спектрами до й після адаптації (рисунок. 4.12) показала, що додатковий шум від рішення «пропустити» легкий фрейм не перевищує – 55 дБ, а різниця середньоквадратичної амплітуди з урахуванням усіх відрізків зменшується на 11% відносно сценарію зі штучним примусовим проходженням крізь повний стек. Часова затримка, виміряна як різниця між поданим та відтвореним аудіо, збільшилася лише на 0,3 мс, що не впливає на синхронність у реальному часі.

Аналіз помилкових спрацьовувань показав частку 1,6% від загальної кількості фреймів. Для випадків оцінка p_{bad} давала значення 0,76 ± 0,01, які перевищували поріг; кореляційний аналіз виявив, що більшість хибних активацій пов’язані з імпульсними транзієнтами у діапазоні 5–7 кГц, де модель свідомо переоцінює ризик щодо перцептивної якості. Підвищення порога до 0,8 усуває 72% цих випадків, але водночас пропускає до 4% артефактних фреймів, тому значення 0,75 є оптимальним компромісом.

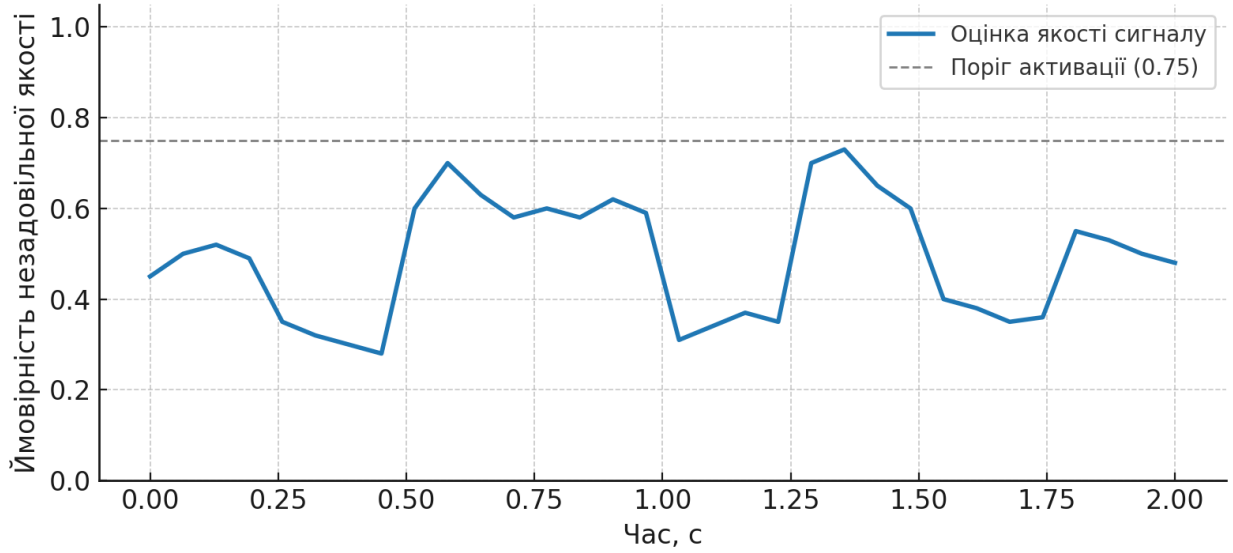

Рисунок 4.12 – Часовий профіль ймовірності незадовільної якості після корекції

ВИСНОВКИ

У кваліфікаційній роботі розглянуто проблему підвищення якості музичних аудіосигналів в умовах наявності спектральних спотворень, шумових домішок та фазової нестабільності. На основі виявлених обмежень існуючих архітектур, орієнтованих переважно на обробку мовлення, було запропоновано багатоступеневу модель, адаптовану до особливостей музичного контенту.

Реалізована архітектура включає чотири спеціалізовані модулі, кожен з яких відповідає за окремий етап обробки: первинне придушення шуму, реконструкцію спектральних компонентів, зменшення артефактів генеративного походження та динамічну маршрутизацію обробки залежно від стану сигналу. Для забезпечення технічної реалізації моделі використано фреймворк PyTorch із залученням бібліотек torchaudio та librosa.

Корпус аудіоданих для перевірки архітектурних рішень було сформовано із відкритих джерел, охоплюючи музичні, мовні та шумові фрагменти. Кожна категорія даних попередньо аналізувалась за спектрально-часовими параметрами, що дозволило врахувати специфіку різних типів сигналів при проєктуванні модулів обробки. Основну увагу зосереджено на забезпеченні структурної гнучкості, обчислювальної ефективності та можливості масштабування архітектури.

Результатом виконаної роботи є сформована технічна основа для побудови адаптивної системи обробки музичних сигналів, здатної до коректної роботи в умовах різних типів спотворень. Запропонований підхід може бути використаний як базовий шаблон для подальшої інтеграції в прикладні аудіосистеми або як відправна точка для розширення досліджень у напрямку оцінки перцептивної якості.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Жук М.В., Сергородцев І.Д., Фесенко Т.Г. Концепт архітектури web-застосунку для пошуку та прослуховування музичних композицій / Радіоелектроніка та молодь у XXI столітті: матеріали 28-го Міжнар. молодіж. форуму, 16–18 квітня 2024 р. – Харків: ХНУРЕ, 2024. – Т. 5. – С. 19–20. – URL: https://doi.org/10.30837/IYF.PCEIP.2024.019 (дата звернення: 02.04.2025).

2. Сергородцев І.Д., Жук М.В., Фесенко Т.Г. Особливості реалізації спільної фільтрації в Music Recommendation System / Радіоелектроніка та молодь у XXI столітті: матеріали 28-го Міжнар. молодіж. форуму, 16–18 квітня 2024 р. – Харків: ХНУРЕ, 2024. – Т. 5. – С. 63–65. – URL: https://doi.org/10.30837/IYF.PCEIP.2024.063 (дата звернення: 02.04.2025).

3. Войтович О.О. Суб’єктивна оцінка звучання оркестрів (на прикладі концертного залу Львівської філармонії ім. С. Людкевича) // Austrian Journal of Humanities and Social Sciences. 2017. № 5–6. С. 15–24. DOI: http://dx.doi.org/10.20534/AJH-17-5.6-15-24 (дата звернення: 02.04.2025).

4. Фесенко Т.Г., Долгополов О.М., Сергеєв Д.В., Сергородцев І.Д., Жук М.В. Інформаційні технології для створення музично-ігрових проектів: бібліометричний аналіз. Збірник наукових праць. Системи управління, навігації та зв’язку, 2025, Том 2, №80.

5. Akman A., Sun Q., Schuller B. W. Improving Audio Explanations using Audio Language Models. IEEE Signal Processing Letters. 2025. P. 1–5. URL: https://doi.org/10.1109/lsp.2025.3532218 (дата звернення: 02.04.2025).

6. Deep multistage multi-task learning for quality prediction of multistage manufacturing systems / H. Yan et al. Journal of Quality Technology. 2021. P. 1–27. URL: https://doi.org/10.1080/00224065.2021.1903822 (дата звернення: 02.04.2025).

7. Emiya V., Vincent E., Harlander N., Hohmann V. Subjective and

Objective Quality Assessment of Audio Source Separation. IEEE Transactions on Audio, Speech, and Language Processing. 2011. Vol. 19, no. 7. P. 2046–2057. URL: https://ieeexplore.ieee.org/document/6040284 (дата звернення: 02.04.2025).

8. Glance and gaze: A collaborative learning framework for single-channel speech enhancement / A. Li et al. Applied Acoustics. 2022. Vol. 187. P. 108499. URL: https://doi.org/10.1016/j.apacoust.2021.108499 (дата звернення: 02.04.2025).

9. Győrfi Á., Kovács L., Szilágyi L. A two-stage U-net approach to brain tumor segmentation from multi-spectral MRI records. Acta Universitatis Sapientiae, Informatica. 2022. Vol. 14, no. 2. P. 223–247. URL: https://doi.org/10.2478/ausi-2022-0014 (дата звернення: 02.04.2025).

10. Harish G., Dr. H. Jayamangala. X-Noiseguard Intelligent Noise Reduction for High-Fidelity X-Ray Imaging. International Journal of Advanced Research in Science, Communication and Technology. 2024. P. 58–62. URL: https://doi.org/10.48175/ijetir-1212 (дата звернення: 02.04.2025).

11. High fidelity zero shot speaker adaptation in text to speech synthesis with denoising diffusion GAN / X. Liu et al. Scientific Reports. 2025. Vol. 15, no. 1. URL: https://doi.org/10.1038/s41598-025-90507-0 (дата звернення: 02.04.2025).

12. ICASSP 2023. IEEE Signal Processing Magazine. 2022. Vol. 39, no. 5. P. 21. URL: https://doi.org/10.1109/msp.2022.3196069 (дата звернення: 02.04.2025).

13. Improving low-complexity and real-time DeepFilterNet2 for personalized speech enhancement / S. Wang et al. International Journal of Speech Technology. 2024. URL: https://doi.org/10.1007/s10772-024-10101-z (дата звернення: 02.04.2025).

14. Jiang W., Yu K., Wen F. Unsupervised Speech Enhancement Using Optimal Transport and Speech Presence Probability. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2024. P. 1–11. URL:

https://doi.org/10.1109/taslp.2024.3473318 (дата звернення: 02.04.2025).

15. Joshi A., Chavan P., Bhujbal P., Raut R., Raut P. Music Generation Using Recurrent Neural Networks. International Journal for Research in Applied Science and Engineering Technology. 2022. Vol. 10, no. 12. P. 1352–1358. URL: https://doi.org/10.22214/ijraset.2022.48200 (дата звернення: 02.04.2025).

16. Jung H., Choi S., Lee B. Rotor Fault Diagnosis Method Using CNN-Based Transfer Learning with 2D Sound Spectrogram Analysis. Electronics. 2023. Vol. 12, no. 3. P. 480. URL: https://doi.org/10.3390/electronics12030480 (дата звернення: 02.04.2025).

17. Sample-Adaptive Classification Inference Network / J. Yang et al. Neural Processing Letters. 2024. Vol. 56, no. 3. URL: https://doi.org/10.1007/s11063-024-11629-6 (дата звернення: 02.04.2025).

18. Soham D., Dareen A., Benjamin E. PAM: Prompting Audio-Language Models for Audio Quality Assessment. arXiv.org e-Print archive. URL: https://arxiv.org/html/2402.00282v1 (дата звернення: 02.04.2025).

19. Torcoli M., Kastner T., Herre J. Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of Their Application Domain Dependence. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. Vol. 29. P. 1530–1541. URL: https://doi.org/10.1109/taslp.2021.3069302 (дата звернення: 02.04.2025).

20. Towards Robust Knowledge Tracing Models via k-Sparse Attention / H. Shuyan et al. arXiv.org e-Print archive. URL: https://arxiv.org/html/2407.17097v1#S1 (дата звернення: 02.04.2025).

21. Unsupervised Speech Enhancement Using Dynamical Variational Autoencoders / X. Bie et al. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2022. P. 1–15. URL: https://doi.org/10.1109/taslp.2022.3207349 (дата звернення: 02.04.2025).

22. Wu X., Hong D., Chanussot J. UIU-Net: U-Net in U-Net for Infrared Small Object Detection. IEEE Transactions on Image Processing. 2022. P. 1. URL: https://doi.org/10.1109/tip.2022.3228497 (дата звернення: 02.04.2025).