**ПОЛТАВСЬКИЙ ДЕРЖАВНИЙ АГРАРНИЙ УНІВЕРСИТЕТ**
**Навчально-науковий інститут економіки, управління, права та інформаційних технологій**
**Кафедра інформаційних систем та технологій**

# КВАЛІФІКАЦІЙНА РОБОТА

на здобуття ступеня вищої освіти магістр

на тему: **«Методологія синтезу мовлення з використанням технології нейронних мереж»**

Виконав: здобувач вищої освіти
за освітньою програмою
Інформаційні управляючі системи та
технології спеціальності
126 Інформаційні системи та технології
ступеня вищої освіти магістр
групи 126ІСТ_мд_2024
Єфремов Андрій Валерійович
Керівник: Протас Надія Михайлівна
Рецензент: Ковальчук Станіслав Богданович

**Полтава – 2025 року**

**ПОЛТАВСЬКИЙ ДЕРЖАВНИЙ АГРАРНИЙ УНІВЕРСИТЕТ**
**Навчально-науковий інститут економіки, управління, права та інформаційних технологій**
**Кафедра інформаційних систем та технологій**

Освітня програма Інформаційні управляючі системи та технології
Спеціальність 126 Інформаційні системи та технології
Рівень вищої освіти другий (магістерський)

**ЗАТВЕРДЖУЮ**
Завідувач кафедри
_____ Юрій УТКІН

«08» листопада 2024 року

# З А В Д А Н Н Я
# НА КВАЛІФІКАЦІЙНУ РОБОТУ ЗДОБУВАЧА ВИЩОЇ ОСВІТИ

**Єфремова Андрія Валерійовича**

1. Тема кваліфікаційної роботи: «Методологія синтезу мовлення з використанням технології нейронних мереж».

Керівник роботи к.с.-г.н., доцент, доцент кафедри інформаційних систем та технологій Протас Надія Михайлівна.

Затверджено наказом закладу вищої освіти від від «31» жовтня 2025 року № 1332-ст

2. Строк подання здобувачем вищої освіти роботи «09» грудня 2025 р.

3. Вихідні дані до роботи: науково-технічна література; наукові статті та публікації, доступні в наукометричних базах даних Google Scholar, IEEE Xplore, SpringerLink, ScienceDirect, Scopus, ResearchGate; вітчизняні та міжнародні стандарти та методології; технічна документація до нейромережевих моделей Tacotron 2, FastSpeech 2, WaveNet, VITS, Glow-TTS, а також офіційна документація бібліотек PyTorch, TensorFlow, ESPnet, Coqui TTS і Hugging Face Transformers; інтернет-джерела аналітичного характеру за темою роботи, блоги, статті та технічні звіти експертів у галузі синтезу мовлення.

4. Зміст пояснювальної записки (перелік питань, які потрібно розробити):

Розділ 1. Теоретичні засади синтезу мовлення

Розділ 2. Проєктування та програмна реалізація системи синтезу мовлення

Розділ 3. Експериментальне дослідження ефективності системи

5. Перелік графічного матеріалу: схеми, рисунки, діаграми за темою та об’єктом дослідження.

6. Консультанти розділів кваліфікаційної роботи

| Розділ | Прізвище, ініціали та посада консультанта | Підпис, дата | |
|---|---|---|---|
| | | завдання видав | завдання отримав |
| Оцінювання економічної ефективності результатів дослідження | Калініченко О. В., к. е. н., доцент, доцент кафедри економіки та публічного управління | 24.11.2025 | 04.12.2025 |

7. Дата видачі завдання «08» листопада 2024 р.

КАЛЕНДАРНИЙ ПЛАН

| № з/п | Назва етапів роботи | Строк виконання етапів кваліфікаційної роботи | Примітка |
|---|---|---|---|
| 1. | Вибір і затвердження теми роботи | 29.10.2024 р. | |
| 2. | Складання і затвердження розгорнутого плану та завдання на кваліфікаційну роботу | 30.10.2024 р. – 08.11.2024 р. | |
| 3. | Опрацювання джерел інформації | 11.11.2024 р. – 27.12.2024 р. | |
| 4. | Збір, вивчення і обробка інформації, необхідної для виконання роботи | 30.12.2024 р.– 19.01.2025 р. | |
| 5. | Виконання теоретико-методологічного розділу роботи | 17.02.2025 р.– 16.05.2025 р. | |
| 6. | Виконання дослідницько-аналітичного розділу роботи | 02.06.2025 р.– 13.07.2025 р. | |
| 7. | Виконання проєктно-рекомендаційного розділу роботи | 08.09.2025 р.– 14.11.2025 р. | |
| 8. | Оцінювання економічної ефективності результатів дослідження | 24.11.2025 р.– 04.12.2025 р. | |
| 9. | Оформлення тексту роботи | 05.12.2025 р.– 08.12.2025 р. | |
| 10. | Попередній захист роботи на кафедрі | 09.12.2025 р. | |
| 11. | Доопрацювання роботи з урахуванням зауважень і пропозицій | 10.12.2025 р.- 14.12.2025 р. | |
| 12. | Нормоконтроль | 15.12.2025 р. – 16.12.2025 р. | |
| 13. | Захист кваліфікаційної роботи | 18.12.2025 р. | |

**Здобувач вищої освіти** **Андрій ЄФРЕМОВ**

**Керівник роботи** **Надія ПРОТАС**

**ПОЛТАВСЬКИЙ ДЕРЖАВНИЙ АГРАРНИЙ УНІВЕРСИТЕТ**
**НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ЕКОНОМІКИ, УПРАВЛІННЯ, ПРАВА ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**
**КАФЕДРА ІНФОРМАЦІЙНИХ СИСТЕМ ТА ТЕХНОЛОГІЙ**

ЄФРЕМОВ АНДРІЙ ВАЛЕРІЙОВИЧ

# «МЕТОДОЛОГІЯ СИНТЕЗУ МОВЛЕННЯ З ВИКОРИСТАННЯМ ТЕХНОЛОГІЇ НЕЙРОННИХ МЕРЕЖ»

Освітньо-професійна програма
Інформаційні управляючі системи та технології
Спеціальність 126 Інформаційні системи та технології
Ступінь вищої освіти Магістр

РЕФЕРАТ
кваліфікаційної роботи на здобуття кваліфікації –
магістр з інформаційних систем та технологій

Полтава – 2025 року

Кваліфікаційна робота складається зі вступу, трьох розділів, висновків, списку використаних джерел і додатків. Основний текст викладено на 70 сторінках, містить 21 рисунок і 16 таблиць. Список використаних джерел налічує 63 найменування.

## Основний зміст роботи

Теоретична частина роботи присвячена аналізу сучасних підходів до синтезу мовлення та ролі нейронних мереж у побудові високоякісних TTS-систем. Розглянуто базові принципи синтезу мовлення, структуру двоступеневого пайплайну «лінгвістичне кодування – акустична модель – вокодер», охарактеризовано етапи нормалізації тексту, фонемізації, узгодження тексту з акустичними ознаками. Систематизовано методи акустичної репрезентації (мел-спектрограми, контур $F_0$, енергія, озвученість), а також наведено класифікацію методів синтезу мовлення – конкатенативного, формантного та нейромережевого. Окрему увагу приділено огляду нейромережевих архітектур Tacotron, FastSpeech, VITS, їхнім сильним сторонам, обмеженням та сферам доцільного застосування, а також порівнянню сучасних вокодерів WaveNet і HiFi-GAN з погляду якості звучання та швидкодії.

Практична частина роботи демонструє побудову методології синтезу мовлення з використанням глибоких нейронних мереж. Розроблено архітектуру системи синтезу мовлення, що включає модуль лінгвістичного препроцесингу, акустичну модель на основі сучасної нейромережевої архітектури та нейронний вокодер. Обґрунтовано вибір моделі синтезу (Tacotron-подібний або FastSpeech-подібний підхід із предиктором тривалості) з урахуванням вимог до швидкості інференсу й керованості просодикою. Описано процедури формування й очищення навчального корпусу, нормалізації аудіоданих, налаштування гіперпараметрів і функцій втрат, а також реалізацію механізму переозвучення аудіофайлів з використанням спікер-залежних ембеддингів. Здійснено програмну реалізацію модулів системи із застосуванням сучасних бібліотек глибокого навчання, забезпечено можливість зміни тембру, темпу та інтонаційних характеристик синтезованого мовлення.

Проєктно-рекомендаційний розділ містить результати експериментальної оцінки якості синтезу мовлення, аналіз отриманих метрик та обґрунтування економічної доцільності впровадження розробленої системи. Проведено організацію модельного експерименту з використанням суб’єктивних (MOS-оцінювання) та об’єктивних показників (PESQ, STOI, MCD, показники швидкодії). Порівняно результати розробленої системи з базовими або відкритими рішеннями, визначено переваги обраної архітектури за природністю звучання, стабільністю вирівнювання та гнучкістю керування просодикою. Виконано техніко-економічне обґрунтування використання нейромережевого синтезу мовлення в прикладних сервісах (голосові інтерфейси, освітні платформи, інклюзивні системи), показано можливість зниження витрат на озвучення контенту та підвищення доступності інформаційних послуг за рахунок автоматизації генерації мовлення.

## Висновки

Метою дослідження було розробити та обґрунтувати методологію синтезу мовлення з використанням технології нейронних мереж, яка забезпечує високий рівень природності звучання, керованість просодикою та придатність до інтеграції в сучасні інформаційні системи. Запропонований підхід дозволяє поєднати переваги глибоких нейронних архітектур із чітко структурованим пайплайном обробки тексту й акустичних даних.

Результати аналізу сучасних методів синтезу мовлення підтвердили доцільність застосування нейромережевих моделей як базового інструменту для генерації мовлення, наближеного до природного. Порівняння конкатенативних, формантних та нейромережевих підходів показало, що саме глибокі моделі на кшталт Tacotron, FastSpeech, VITS забезпечують оптимальний баланс між якістю, масштабованістю та можливостями персоналізації голосу.

У роботі розроблено архітектуру системи синтезу мовлення, що включає модуль текстової нормалізації й лінгвістичного кодування, акустичну модель та нейронний вокодер. Сформовано методичні рекомендації щодо підготовки навчального корпусу, вибору частоти дискретизації, налаштування функцій втрат і регуляризації, що дає змогу забезпечити стабільність навчання та відтворюваність результатів. Запропоновано підхід до реалізації переозвучення аудіоматеріалів із використанням спікер-орієнтованих ембеддингів, що розширює можливості персоналізації голосових сервісів.

Експериментальна оцінка розробленої системи за суб'єктивними та об'єктивними показниками продемонструвала високу природність синтезованого мовлення, прийнятну швидкодію та відсутність суттєвих артефактів у широкому діапазоні тестових сценаріїв. Порівняння з базовими рішеннями підтвердило покращення якості звучання та гнучкості керування просодикою за збереження придатності до роботи у режимі, близькому до реального часу.

Техніко-економічне обґрунтування впровадження розробленої методології показало перспективність її використання в інформаційних системах, орієнтованих на масову генерацію голосового контенту. Зменшення витрат на запис та обробку мовного матеріалу, а також можливість швидкої адаптації голосових моделей під нові домени та мовців створюють передумови для підвищення ефективності цифрових сервісів.

Наукова новизна роботи полягає у систематизації нейромережевих підходів до синтезу мовлення, формуванні узагальненої методології побудови TTS-систем на основі глибоких нейронних мереж та експериментальному обґрунтуванні впливу архітектурних рішень на якість і швидкодію синтезованого мовлення. Практичний результат полягає у створенні методичних та архітектурних основ для розробки прикладних систем синтезу мовлення, придатних до інтеграції в комерційні й дослідницькі програмні продукти.

## Список публікацій здобувача

1. Єфремов А. Еволюція комп'ютерного синтезу мовлення і роль глибинного навчання. Матеріали *науково-практичної конференції за підсумками проходження виробничих практик здобувачів вищої освіти спеціальності 126 Інформаційні системи та технології, кафедра інформаційних систем та технологій Полтавського державного аграрного університету, 22 жовтня 2025 р.* Вип. XI. Полтава: ПДАУ, 79 с.

2. Єфремов А. Підготовка навчального корпусу для моделі синтезу мовлення. *Студентські роботи за науковою тематикою кафедри інформаційних систем та технологій: матеріали XXII щорічного міждисциплінарного семінару, 25 листопада 2025 р.* Полтава: ПДАУ, 2025 р.

3. Флегантов Л., Єфремов А. Комплексний аналіз методологій оцінки інтелектуальних систем синтезу та розпізнавання мовлення. *Progressive Approaches in Science and Engineering: Collection of Scientific Papers with Proceedings of the 2nd International Scientific and Practical Conference.* International Scientific Unity. November 26-28, 2025. Copenhagen, Denmark. 283-289 p.

## АНОТАЦІЯ

Єфремов А.В. «Методологія синтезу мовлення з використанням технології нейронних мереж». Кваліфікаційна робота на правах рукопису.

Кваліфікаційна робота на здобуття ступеня вищої освіти магістр за освітньо-професійною програмою «Інформаційні управляючі системи та технології» спеціальності 126 «Інформаційні системи та технології». Полтавський державний аграрний університет, Полтава, 2025.

Робота присвячена проблематиці побудови сучасних систем синтезу мовлення на основі нейронних мереж. Досліджено теоретичні та прикладні аспекти процесу синтезу мовлення, проаналізовано еволюцію методів від конкатенативних і формантних підходів до глибоких нейромережевих моделей. Визначено вимоги до

лінгвістичного препроцесингу, акустичної репрезентації та вокодерів, що забезпечують високу якість, масштабованість і керованість синтезованого мовлення.

У першому розділі виконано огляд принципів синтезу мовлення, класифіковано сучасні підходи та розглянуто нейромережеві архітектури Tacotron, FastSpeech, VITS, а також вокодери WaveNet і HiFi-GAN. Показано, як вибір архітектури впливає на природність звучання, швидкодію та стабільність системи. У другому розділі розроблено архітектуру системи синтезу мовлення з використанням глибоких нейронних мереж, описано підготовку корпусу, процедури навчання, механізм переозвучення аудіофайлів та засоби керування голосовими характеристиками. Третій розділ присвячено експериментальній перевірці запропонованої методології, оцінюванню якості синтезованого мовлення за суб'єктивними та об'єктивними метриками та техніко-економічному обґрунтуванню впровадження системи у практичні інформаційні сервіси.

Робота має значення для розроблення голосових інтерфейсів, мультимедійних і освітніх платформ, інклюзивних технологій та інших систем, де потрібне автоматичне генерування природного мовлення. Запропоновані підходи можуть бути використані для подальшої оптимізації архітектур нейромережевого синтезу мовлення, підвищення якості звуку та розширення можливостей персоналізації голосу.

Ключові слова: синтез мовлення, нейронні мережі, Tacotron, FastSpeech, VITS, WaveNet, HiFi-GAN, вокодер.

## ANNOTATION

Yefremov A.V. «Methodology of speech synthesis using neural network technology». Master's thesis manuscript.

The master's thesis for obtaining a Master's degree in the educational and professional program «Information Management Systems and Technologies», specialty 126 «Information Systems and Technologies». Poltava State Agrarian University, Poltava, 2025.

The thesis is devoted to the development of modern speech synthesis systems based on neural networks. Theoretical and applied aspects of text-to-speech (TTS) generation are investigated, including the evolution from concatenative and formant methods to deep neural architectures. The work identifies the requirements for linguistic preprocessing, acoustic representations and vocoders that ensure high quality, scalability and controllability of synthesized speech.

The first section analyzes the principles of speech synthesis and classifies current approaches, with a detailed review of neural models such as Tacotron, FastSpeech and VITS, as well as vocoders WaveNet and HiFi-GAN. The influence of architectural choices on naturalness, inference speed and system stability is examined. The second section presents the architecture of a speech synthesis system based on deep neural networks, including text normalization, corpus preparation, model training, implementation of re-voicing mechanisms and control of voice characteristics. The third section is focused on the experimental evaluation of the proposed methodology, combining subjective and objective quality metrics (MOS, PESQ, STOI, MCD) and providing an economic feasibility study of integrating the system into practical information services.

results of the thesis are relevant for the development of voice interfaces, multimedia and educational platforms, inclusive technologies and other applications where automatic generation of natural-sounding speech is required. The proposed methodology can be used for further optimization of neural TTS architectures, improving audio quality and extending personalization capabilities of synthetic voices.

Keywords: speech synthesis, neural networks, Tacotron, FastSpeech, VITS, WaveNet, HiFi-GAN, vocoder.

# ЗМІСТ

ВСТУП ..... 6
РОЗДІЛ 1. ТЕОРЕТИЧНІ ЗАСАДИ СИНТЕЗУ МОВЛЕННЯ ..... 10
1.1 Принципи організації систем синтезу мовлення ..... 10
1.2 Еволюція методів синтезу мовлення ..... 12
1.3 Архітектури нейромережевих моделей для синтезу мовлення ..... 15
1.4 Нейронні вокодери та технології реконструкції звукового сигналу ..... 16
1.5 Методи перетворення та стилізації голосових характеристик ..... 19
Висновки до розділу 1 ..... 23
РОЗДІЛ 2. ПРОЄКТУВАННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ СИСТЕМИ СИНТЕЗУ МОВЛЕННЯ ..... 25
2.1 Обґрунтування вибору архітектури нейронної мережі ..... 25
2.2 Формування та препроцесинг навчального корпусу для моделі синтезу мовлення ..... 26
2.3 Алгоритмічна реалізація процесу навчання моделі ..... 28
2.4 Інтеграція підсистеми переозвучення аудіофалів ..... 36
2.5 Оптимізація моделі для підвищення якості мовлення ..... 39
Висновки до розділу 2 ..... 41
РОЗДІЛ 3. ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ СИСТЕМИ ..... 43
3.1 Організація експерименту та методологія тестування ..... 43
3.2 Методика оцінювання якості синтезованого мовлення ..... 47
3.3 Порівняльний аналіз отриманих результатів із базовими моделями ..... 53
3.4 Аналіз артефактів генерації та напрями вдосконалення моделі синтезу мовлення ..... 58
3.5 Оцінка економічної ефективності синтезу мовлення ..... 64
Висновки до розділу 3 ..... 71
ВИСНОВКИ ..... 73
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ ..... 75
ДОДАТКИ ..... 80

# ВСТУП

Технології синтезу мовлення є надзвичайно важливим елементом сучасних людино-машинних інтерфейсів. Вони знаходять широке застосування у віртуальних асистентах, навігаційних системах та інклюзивних технологіях. На відміну від традиційних методів синтезу мовлення (конкатенативного, формантного), які мають обмежену виразність та адаптивність, сучасні нейромережеві архітектури (Tacotron, FastSpeech, HiFi-GAN) дозволяють моделювати складні нелінійні залежності. Це забезпечило якісний прорив у галузі, наблизивши звучання синтезованого голосу до природного людського мовлення

*Актуальність теми* зумовлена стрімким розвитком технологій штучного інтелекту та необхідністю створення високоякісних, гнучких і адаптивних систем синтезу мовлення, здатних працювати в режимі реального часу, підтримувати багатомовність і враховувати емоційні та інтонаційні особливості мовлення.

Розробка методології синтезу мовлення з використанням нейронних мереж є важливим завданням для підвищення ефективності комунікаційних технологій, розширення можливостей доступності інформації та розвитку людино орієнтованих цифрових сервісів.

*Зв'язок роботи з науковими програмами, планами, темами*. Магістерська кваліфікаційна робота безпосередньо відповідає дослідженням у межах науково-дослідної ініціативної тематики «Організаційно-методологічні аспекти впровадження інформаційно-комунікаційних систем і технологій в управлінні діяльністю сучасних організацій та підприємств за умов переходу до цифрової економіки» (ДРН 0123U105060, 2023–2028 рр.), що реалізується на кафедрі інформаційних систем та технологій, а також узгоджується з тематикою досліджень навчально-дослідної лабораторії інтелектуальних систем, комп'ютерних мереж і інтернету речей кафедри інформаційних систем та технологій Полтавського державного аграрного університету.

*Мета роботи* полягає у розробленні та обґрунтуванні методології синтезу мовлення з використанням технології нейронних мереж.

*Завдання роботи*:

- дослідити теоретико-методологічні основи синтезу мовлення та сучасні тенденції його розвитку;
- проаналізувати архітектурні принципи та можливості традиційних систем синтезу мовлення;
- вивчити сучасні нейромережеві моделі для синтезу мовлення (Tacotron, WaveNet, FastSpeech, VITS);
- розробити методологічну схему побудови системи синтезу мовлення на основі глибоких нейронних мереж;
- здійснити порівняльний аналіз ефективності традиційних і нейромережевих методів синтезу;
- сформувати рекомендації щодо впровадження нейромережевих технологій у практичні системи синтезу мовлення.

*Об'єкт дослідження* – процес синтезу мовлення.

*Предмет дослідження* – методи та архітектури нейронних мереж, що застосовуються для синтезу мовлення.

*Методи дослідження*:

- аналіз наукових і технічних джерел з проблєматики синтезу мовлення та нейронних мереж;
- системний аналіз архітектур сучасних моделей глибокого навчання;
- експериментальне моделювання процесу синтезу мовлення на основі нейромережевих технологій;
- порівняльна оцінка якості синтезу за метриками MOS, PESQ, STOI, швидкістю генерації та складністю обчислень;
- узагальнення отриманих результатів для формування рекомендацій з підвищення ефективності систем синтезу мовлення.

*Інформаційна база дослідження* включає наукові статті, монографії, технічну документацію до моделей Tacotron 2, FastSpeech 2, WaveNet, VITS,

Glow-TTS, матеріали з наукометричних баз Scopus, IEEE Xplore, SpringerLink, а також офіційну документацію бібліотек PyTorch, TensorFlow, ESPnet, Coqui TTS і Hugging Face Transformers.

*Елементи наукової новизни:*

- систематизація сучасних нейромережевих підходів до синтезу мовлення;
- обґрунтування критеріїв вибору архітектури моделі для забезпечення балансу між якістю звучання та швидкодією;
- розроблення узагальненої методології побудови системи синтезу мовлення з використанням глибоких нейронних мереж;
- експериментальна оцінка впливу архітектурних параметрів моделі на якість синтезованого мовлення.

*Практична значущість* отриманих результатів полягає в можливості їх застосування під час створення комерційних і дослідницьких систем синтезу мовлення для мультимедійних сервісів, мобільних застосунків, автоматизованих довідкових систем, освітніх платформ та інклюзивних технологій. Розроблена методологія може бути використана для побудови адаптивних голосових систем, здатних генерувати індивідуалізоване мовлення з високим ступенем природності.

*Апробація результатів дослідження.* За результатами дослідження опубліковано тези доповідей: «Еволюція комп'ютерного синтезу мовлення і роль глибинного навчання». Матеріали науково-практичної конференції за підсумками проходження виробничих практик здобувачів вищої освіти спеціальності 126 Інформаційні системи та технології, кафедра інформаційних систем та технологій Полтавського державного аграрного університету, 22 жовтня 2025 р. Вип. XI. Полтава: ПДАУ, 2025; «Підготовка навчального корпусу для моделі синтезу мовлення». Студентські роботи за науковою тематикою кафедри інформаційних систем та технологій: матеріали XXII щорічного міждисциплінарного семінару, 25 листопада 2025 р. Полтава: ПДАУ, 2025; «Комплексний аналіз методологій оцінки інтелектуальних систем

синтезу та розпізнавання мовлення. Progressive Approaches in Science and Engineering: Collection of Scientific Papers with Proceedings of the 2nd International Scientific and Practical Conference. International Scientific Unity. November 26-28, 2025. Copenhagen, Denmark, 2025.

*Структура та обсяг кваліфікаційної роботи.* Магістерська робота складається зі вступу, трьох розділів, висновків, списку використаних джерел і додатків. Основний текст викладено на 70 сторінках, містить 21 рисунок і 16 таблиць. Список використаних джерел налічує 63 найменування.

# РОЗДІЛ 1
# ТЕОРЕТИЧНІ ЗАСАДИ СИНТЕЗУ МОВЛЕННЯ

## 1.1 Принципи організації систем синтезу мовлення

Синтез мовлення – це перетворення тексту на природний звуковий сигнал. Сучасна архітектура синтезу мовлення зазвичай є двоступеневою: спершу генерується проміжне акустичне подання (мел-спектрограма, $F_0$, енергія), після чого нейронний вокодер реконструює часовий сигнал. Такий підхід дозволяє гнучко налаштовувати параметри генерації.

Якість синтезу мовлення значною мірою залежить від лінгвістичного препроцесингу: нормалізації тексту (числа, абревіатури) та його перетворення у фонемну або символьну послідовність. Вибір рівня подання впливає на стабільність навчання та точність відповідності [1].

Узгодження тексту з акустикою реалізується двома методами: через механізм «уваги» (Tacotron-подібні моделі), що забезпечує гнучкість, але схильний до збоїв, або через явний предиктор тривалості (FastSpeech-подібні), що гарантує стабільність і швидкість генерації [2]. Акустичне моделювання базується на мел-спектрограмах, а додавання контурів $F_0$ та міток озвученості дозволяє керувати інтонацією та емоційним забарвленням [3].

За фінальну якість звуку відповідають нейронні вокодери. Авторегресивні моделі (WaveNet) є точними, але повільними, тоді як GAN-вокодери (HiFi-GAN) забезпечують високу якість у реальному часі. Вибір архітектури зумовлений необхідною частотою дискретизації (SR) [4].

Керування просодикою (темп, ритм, інтонація) досягається через латентні коди стилю або предиктори енергії та висоти тону. Навчання оптимізується комбінацією регресійних, перцептивних та адверсаріальних функцій втрат.

Узагальнюваність моделі синтезу мовлення залежить від різноманітності даних (баланс мовців, стилів, фонем) та якості їх препроцесингу (єдина SR,

нормування рівня, очищення шумів) [5]. Стабільність системи підтримується регуляризацією, а валідація включає як суб'єктивні (MOS), так і об'єктивні метрики: спектральну точність (MCD), розпізнавання (WER/CER) та подібність до мовця [6].

Узагальнення основних принципів синтезу мовлення представлене у таблиці 1.1.

Таблиця 1.1 – Узагальнення основних принципів синтезу мовлення

| Принцип | Головне питання | Типові рішення | Наслідки для якості |
|---|---|---|---|
| Лінгвістичне кодування | Як подати текст у зручному для моделі вигляді | Нормалізація, фонемізація або субсловні одиниці | Менше помилок вирівнювання, стійкіша дикція |
| Узгодження текст-аудіо | Як зіставити символи й часові кадри | Навчувана увага (Tacotron) або предиктор тривалості (FastSpeech) | Баланс гнучкості та стабільності, контроль темпу |
| Акустична репрезентація | У чому навчати акустичну модель | Мел-спектрограма + $F_0$, енергія, озвученість(голос/шум) | Кращий контроль, стабільніша інтонація та римт |
| Вокодер | Як відновити хвилю | HiFi-GAN/WaveNet за обраною SR | Природність тембру, відсутність «дзвону»/шумів |
| Просодичний контроль* | Як керувати стилем | Варіаційні латенти, предиктори (енергії/тривалості) | Керовані паузи, темп, емоційні відтінки |
| Дані і препроцесинг | Як забезпечити узагальнюваність | Баланс мовців/стилів, очищення, нормування рівня | Менше артефактів, стабільна вимова |
| Критерії якості | Як оцінювати результат | MOS, ABX-подібність, MCD, WER/CER | Вимірюваність покращень, відтворюваність |

Примітка. Просодичний контроль (просодика) – здатність людини свідомо або несвідомо керувати звучанням своєї мови. Включає керування такими елементами, як: інтонація – підвищення або зниження тону голосу (наприклад, щоб перетворити твердження на запитання); ритм і темп – швидкість мовлення та чергування наголошених і ненаголошених складів; гучність – виділення голосом важливих моментів (логічний наголос); паузи – зупинки для розділення думок або створення ефекту очікування [7].

Структурна схема пайплайну синтезу мовлення включає низку послідовних перетворень, що впливають на результат – лінгвістичне кодування → узгодження текст–аудіо → акустична репрезентація → вокодер; допоміжні

ланки – дані й препроцесинг, просодичний контроль; вихід – критерії якості (MOS, ABX-подібність, MCD, WER/CER), – представлена на рисунку 1.1.

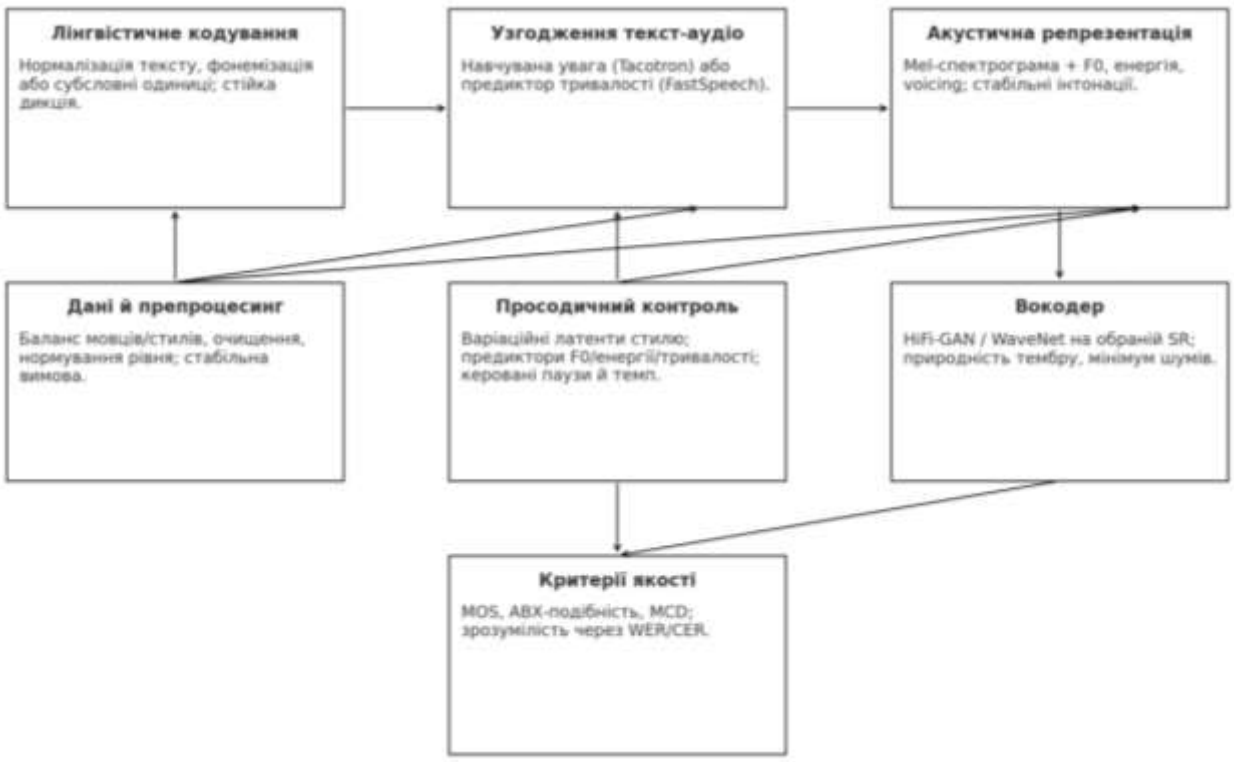


Рисунок 1.1 – Структурна схема пайплайну синтезу мовлення

Отже, практична реалізація синтезу мовлення зводиться до узгодженого вибору лінгвістичного представлення, механізму узгодження, типу акустичної моделі й вокодера, причому якість і керованість синтезу досягаються завдяки адекватним даним, просодичному контролю та коректним метрикам оцінювання.

## 1.2 Еволюція методів синтезу мовлення

Сучасні методи озвучення тексту балансують між природністю звучання, керованістю та вимогами до ресурсів. Це зумовлює постійний пошук компромісів, здатних поєднувати якість синтезу з ефективністю виконання. Виділяють три основні підходи:

- конкатенативний синтез – полягає у комбінуванні («склеюванні») попередньо записаних фрагментів мовлення (від фонем до слів) з великої бази даних. Цей метод забезпечує високу реалістичність тембру, але має суттєві недоліки: залежність якості від обсягу корпусу, складність керування інтонацією, поява артефактів на стиках фрагментів та неможливість зміни стилю без перезапису бази [8];

- формантний синтез – моделює фізичні процеси голосоутворення шляхом керування акустичними параметрами (частота основного тону F0, форманти, шум). Цей підхід є компактним, стійким до будь-якої лексики та повністю керованим, проте результуючий голос звучить «роботизовано» через відсутність природних мікроваріацій;

- нейромережевий синтез – поєднує акустичну модель (перетворення тексту на мел-спектрограму) та нейронний вокодер (відновлення звукової хвилі). Використання механізмів уваги або явних предикторів тривалості та енергії дозволяє досягти високої природності й гнучкої стилізації. Метод ефективно вирішує проблеми масштабування та персоналізації, але вимагає значних обчислювальних ресурсів та ретельної підготовки даних.

Детальне порівняння методів синтезу мовлення представлене у таблиці 1.2.

Таблиця 1.2 – Порівняння методів синтезу мовлення

| Характеристика | Конкатенативний синтез | Формантний синтез | Нейромережевий синтез |
|---|---|---|---|
| Принцип роботи | Поєднує («зшиває») заздалегідь записані фрагменти реального голосу (від фонем до слів). | Відтворює мовлення шляхом моделювання фізичних процесів голосового тракту та резонансів за допомогою параметрів. | Використовує навчені моделі для перетворення тексту в акустичні ознаки, а нейронний вокодер реконструює сигнал. |
| Природність | Висока тембральна правдоподібність (у межах домену), але можливі артефакти на місцях склейки. | Низька («роботизоване» звучання), обмежена виразність мікроваріацій. | Дуже висока, близька до природної, здатна відтворювати складні нюанси. |

*Продовження таблиці 1.2*

| | | | |
|---|---|---|---|
| Керованість (Просодика) | Низька. Складно керувати темпом та інтонацією без погіршення якості. | Висока. Повний контроль над висотою тону, частотами та іншими параметрами. | Висока. Гнучке керування стилем, емоціями, тривалістю та інтонацією. |
| Вимоги до даних/ресурсів | Потребує терабайтів аудіо для покриття всіх контекстів. Велика база даних. | Мінімальні. Компактна реалізація, не потребує великих корпусів даних. | Високі. Вимагає якісних даних, ретельної нормалізації та значних обчислювальних ресурсів для навчання. |
| Масштабованість | Погана. Для зміни голосу чи стилю потрібне повне переозвучення бази. | Складна. Потребує ручної розробки правил для нових мов та стилів. | Відмінна. Підтримує багатомовність, персоналізацію та зміну стилів у межах однієї моделі. |
| Основні переваги | Реалістичність «живого» голосу (за наявності хорошого інвентарю). | Інтерпретованість, стабільність на незнайомій лексиці, прозорість налаштувань. | Баланс між високою якістю та керованістю, відсутність «склейок». |
| Основні недоліки | Помилки на позадоменній лексиці, неможливість гнучкої зміни стилю. | Штучність звучання, складність моделювання природних флуктуацій. | Ресурсна затратність, ризик нестабільності та артефактів у рідкісних контекстах. |

Конкатенативні системи синтезу мовлення у минулому переважали у реалістичності, а формантні – у контролі.

Нейромережевий підхід об’єднує переваги обох підходів, забезпечуючи високу якість та гнучкість налаштування просодики.

Сучасні практики синтезу мовлення поєднують елементи різних традицій: правила текстової нормалізації та фонетичного транскрибування, притаманні формантним підходам, використовуються разом із нейронними акустичними моделями; класичні методики оцінювання зрозумілості, що склалися в епоху конкатенативного синтезу, залишаються валідними для порівняння поколінь систем; а параметричні уявлення про голос і тракт мовця допомагають формувати керовані латенти стилю та «ядерні» ознаки для переозвучення. У підсумку три методи не стільки конкурують, скільки

дозволяють обирати доцільний баланс природності, керованості та інженерної складності для конкретного застосування.

### 1.3 Архітектури нейромережевих моделей для синтезу мовлення

Розвиток технологій синтезу мовлення показує поступовий перехід від послідовних авторегресивних схем до паралельних та повністю інтегрованих архітектур, головними представниками яких є моделі Tacotron, FastSpeech та VITS. Авторегресивна модель Tacotron використовує механізм уваги для генерації спектрограм, що забезпечує виразну інтонацію та природну коартикуляцію, однак цей підхід характеризується повільним інференсом, ризиком збоїв на довгих текстах та потребою в окремому вокодері [9].

Альтернативою є паралельна архітектура FastSpeech, яка замість механізму уваги застосовує явний предиктор тривалості. Це дозволяє прискорити процес генерації у 2–3 рази, гарантувати стабільність темпу та забезпечити керування висотою тону й енергією, хоча якість результату стає залежною від точності попереднього вирівнювання даних та чистоти корпусу [10].

Найбільш сучасним рішенням є інтегрована end-to-end система VITS, що об’єднує акустичну модель і вокодер в єдину мережу, яка навчається спільно за допомогою варіаційного автокодувальника та змагального навчання (GAN). Архітектура VITS забезпечує найвищу природність звучання, автоматичне монотонне вирівнювання та роботу в реальному часі, проте вирізняється складністю навчання та чутливістю до якості даних [11].

Отже, Tacotron є оптимальним для досліджень просодики, FastSpeech – для стабільних виробничих рішень, а VITS забезпечує найкращу якість за умови ретельного налаштування.

Детальне порівняння нейромережевих моделей TTS представлене у таблиці 1.3.

Таблиця 1.3 – Порівняння нейромережевих моделей TTS: Tacotron, FastSpeech, VITS

| Модель | Узгодження текст–аудіо | Характер інференсу | Керованість (тривалість/ F0/енергія) | Типовий вокодер / відновлення хвилі |
|---|---|---|---|---|
| Tacotron (авторегресивна з увагою) | Імпліцитне через attention | Авторегресивний (повільніший) | Опосередкована через увагу та додаткові предиктори | Окремий вокодер (WaveNet, HiFi-GAN) |
| FastSpeech (неавторегресивна) | Явне: предиктор тривалості (вчитель/aligner) | Паралельний (швидкий) | Прямий контроль тривалості, F0, енергії | Окремий вокодер (HiFi-GAN тощо) |
| VITS (енд-ту-енд, GAN+flow) | Нав’язане монотонне вирівнювання в навчанні | Паралельний, близький до real-time | Через латенти/предиктори у спільній моделі | Інтегроване відновлення в єдиній архітектурі |

Вибір між моделями Tacotron, FastSpeech і VITS залежить від пріоритетів проєкту: якщо потрібна максимальна керованість темпом і стабільність, більш практичним буде FastSpeech; якщо важлива інтегрована якість і швидкість без зовнішнього вчителя тривалостей – кращою є VITS; для дослідницьких сценаріїв і тонкої просодики доцільно розглядати Tacotron, за умович ретельної стабілізації уваги. FastSpeech підходить для продуктивних систем, тоді як VITS демонструє найкращий компроміс між якістю й ефективністю. Tacotron лишається цінним інструментом для експериментів і навчальних досліджень у сфері TTS.

## 1.4 Нейронні вокодери та технології реконструкції звукового сигналу

Архітектура сучасних систем синтезу мовлення реалізується як каскад моделей, де завершальним і критично важливим етапом є реконструкція звукової хвилі з параметричних даних. Цю функцію виконує спеціалізований компонент – вокодер, що перетворює проміжне акустичне подання,

згенероване акустичною моделлю (зазвичай мел-спектрограму разом з ознаками $F_0$, енергії та озвученості (сигнал/шум)) на часовий сигнал. Тобто, вокодер – це програма, яка перетворює команди (параметри) на справжній, чутний звук.

Вибір вокодера впливає на природність тембру, рівень шумових артефактів, стабільність на довгих послідовностях і реальний час інференсу. Сучасні TTS-системи опираються на дві провідні парадигми: авторегресивне моделювання сирого звуку (клас WaveNet) та генеративно-змагальні моделі зі спеціальними дискримінаторами для мовлення (клас HiFi-GAN).

Вокодер WaveNet створює звук, передбачаючи гучність (амплітуду) кожного наступного, найдрібнішого шматочка звуку (відліку) на основі всіх попередніх. Для цього він використовує особливий вид обробки даних, який дозволяє моделі «бачити» дуже довгий контекст, не сповільнюючи процес. Інакше, вокодер WaveNet генерує мовлення, моделюючи, з якою ймовірністю з’явиться наступний звуковий сигнал (відлік), ґрунтуючись на всій попередній історії аудіо. Це досягається завдяки використанню розширених згорток (дилатованих) – це як слухати з дедалі більшими інтервалами – що дозволяє моделі швидко «зрозуміти» довгі залежності (інтонацію, ритм) у звуці.[12].

Вокодер HiFi-GAN використовує неавторегресивний підхід, що означає, що він генерує весь звуковий фрагмент одразу, не чекаючи, поки буде створений попередній шматочок звуку. Ця особливість забезпечує надзвичайно високу швидкість, яка є швидшою за реальний час, що усуває затримки. Висока якість цього синтезу, близька до студійної природності, досягається завдяки генератору, який працює спільно з двома спеціалізованими «критиками» (дискримінаторами): один оцінює чистоту та деталі звуку в різних масштабах, а інший – його природний ритм та тональність. Навчання цієї системи поєднує вимоги від цих критиків та необхідність точного відтворення «текстури» голосу, завдяки чому HiFi-GAN ефективно масштабується і добре працює на сучасному обладнанні. Порівняння характеристик розглянутих вокодерів узагальнене у таблиці 1.4.

Таблиця 1.4 – Порівняння вокодерів

| Характеристика | WaveNet | HiFi-GAN |
|---|---|---|
| Клас | Авторегресивна модель. Моделювання сирого звуку. | Неавторегресивна (генеративно-змагальна) модель. Використовує спеціальні дискримінатори. |
| Принцип генерації | Покроковий. Передбачає амплітуду кожного наступного відліку (найдрібнішого шматочка) на основі всієї історії попередніх. | Паралельний. Генерує весь звуковий фрагмент одразу, не чекаючи створення попередніх частин. |
| Механізм роботи | Використовує розширені (дилатовані) згортки для охоплення довгого контексту (інтонації, ритму). | Використовує генератор та два «критика» (дискримінатори): один оцінює деталі звуку, інший – ритм і тональність. |
| Швидкість (Інференс) | Повільніша. Вимагає оптимізованих прискорювачів або дистильованих версій для роботи в реальному часі. | Надзвичайно висока. Працює швидше за реальний час, усуває затримки. |
| Якість | Теоретично краще зберігає тонкі мікроваріації та локальну хвильову структуру. | Забезпечує якість, близьку до студійної, точно відтворює «текстуру» голосу. |
| Контроль артефактів | Забезпечує високу точність, але чутливий до обчислювальних обмежень. | Природно «карає» (пригнічує) дзвін, металізацію та фазові артефакти; уникає надто гладких спектрів. |
| Головний компроміс | Максимальна точність локальної хвильової структури ціною швидкості. | Швидкість і стабільність з можливістю появи «шорсткості» при неправильному балансі налаштувань. |

Вибір технології генерації мовлення визначається оптимальним балансом між точністю відтворення звуку та швидкістю роботи системи. Авторегресивні моделі (типу WaveNet) забезпечують високу деталізацію мікроваріацій звуку, проте потребують значних обчислювальних ресурсів, що ускладнює їх використання в реальному часі. Натомість моделі на базі GAN (HiFi-GAN) ефективно усувають характерні артефакти «металізації» та фазові спотворення, забезпечують високу якість, хоча й потребують точного налаштування ваг функцій втрат для уникнення шорсткості звучання. Важливим фактором стабільності є частота дискретизації: діапазон 22,05–24 кГц гарантує надійну генерацію з мінімумом шумів, тоді як підвищення частоти до 44,1–48 кГц вимагає збільшення ємності моделі та застосування суворішої регуляризації.

Для успішної інтеграції системи важливою є уніфікація попередньої обробки даних (узгодження частоти, формату каналів, рівня гучності та пауз), а також спільна оптимізація вокодера з акустичною моделлю. Необхідно контролювати показник швидкодії (RTF) та стабільність на складних фразах. Для покращення відтворення високих частот рекомендовано застосовувати схеми багатосмугової обробки (PQMF) та багатошкальні функції втрат, а для усунення артефактів на межах фраз («хвостів») – синхронізувати параметри кадрування між модулем екстракції ознак та генератором. У виробничих умовах доцільно використовувати стратегію двох моделей: стабільну «консервативну» версію для довгих текстів та «виразну» – для коротких емоційних реплік, зберігаючи при цьому повну сумісність вхідних акустичних ознак.

## 1.5 Методи перетворення та стилізації голосових характеристик

Зміна голосових характеристик (Voice Conversion, VC) спрямована на перетворення джерела мовлення так, щоб зберегти зміст мовлення і його ритмо-мелодичний каркас, одночасно наблизивши тембр та інші ідентифікаційні ознаки до цільового мовця. Класичний підхід передбачає розділення простору ознак на компоненти «контент» і «спікер», подальшу трансформацію спікер-залежної частини та відновлення часової хвилі за допомогою вокодера. За типом даних розрізняють паралельні підходи, де для навчання використовуються пари однакових висловлювань від джерела й цілі, та непаралельні, які обходяться без вирівняних пар; за кількістю прикладів виділяють багато-прикладні (many-shot), малоприкладні (few-/one-shot) і нульовоприкладні (zero-shot) режими.

Головна ідея сучасних систем синтезу мовлення полягає в тому, щоб явно або неявно «очищати» контент від тембральних та інших акустичних особливостей конкретного мовця. Для цього застосовують інваріантні до

спікера представлення, такі як PPG/BN-ознаки з автоматичного розпізнавання мовлення чи контентні ембеддинги самонавчальних моделей типу HuBERT/Wav2Vec2; у поєднанні з окремим кодувальником спікера (x-vector, d-vector) це дозволяє переносити голос на коротких або відсутніх зразках цільового мовця та працювати у zero-shot режимі [13]. Додатковим елементом є моделювання $F_0$ і енергії, оскільки саме вони визначають інтонаційну стилістику; перетворення контуру висоти виконується або предиктором, або через нормалізуючі перетворення у лог-просторі, що знижує артефакти «гелієвості» та «гавкоту».

Виділяють кілька еволюційних хвиль в розвитку архітектури нейромережевих моделей синтезу мовлення. Ранні статистичні підходи на основі GMM/DNN виконували регресію зіставлених ознак між джерельним і цільовим просторами, покладаючись на паралельні дані та DTW-вирівнювання. Далі з’явилися автоенкодерні схеми з вузьким «вмістовим» латентом (AutoVC), які навчають реконструкцію мовлення за умовою на спікер-ембеддинг і завдяки вузькій «горлу» пригнічують спікер-інформацію; варіаційні та флоу-моделі (VAE/Glow/Flow-VC) роблять це імовірнісно, забезпечуючи гладкий латентний простір і оборотність перетворень. Ган-підходи (CycleGAN-VC, StarGAN-VC, GAN-based VC) додають циклові та доменно-специфічні втрати, що дозволяє навчатися без паралельних пар і краще узгоджувати тембр та артикуляційні мікроваріації.

Головною проблємою VC є керування просодикою, оскільки просте «накладання» тембру не гарантує природної стилізації синтезованого голосу. Ефективна стратегія полягає у розв’язанні завдання на два шари: по-перше, збереження або регресія контурів $F_0$, енергії та тривалостей на рівні фонем чи слів; по-друге, перенесення високорівневих стилістичних латентів (емоційність, напруженість, темп) через варіаційні коди або спеціальні предиктори: це означає, що система не просто копіює звук, а намагається скопіювати емоційний стан і манеру мовлення, закодовані у вигляді прихованих числових параметрів.

Порівняння підходів до зміни голосових характеристик за представленнями та навчальними сигналами узагальнене у таблиці 1.5.

Таблиця 1.5 – Порівняння методів зміни голосових характеристик

| Клас підходу | Тип представлення контенту | Дані/вирівнювання | Передача просодики | Типова швидкодія інференсу |
|---|---|---|---|---|
| GMM/DNN-регресія (паралельна) | MFCC/мел кадри (джерело) | Потрібні паралельні пари + DTW | Обмежена, через пост-обробку F0 | Низька–середня |
| AutoVC/AE-схеми | Вузький латент (bottleneck) | Непаралельні, достатньо багатомовців | Базова, через окремий F0-трек | Середня–висока |
| VAE/Flow-VC | Імовірнісний/оборотний латент | Непаралельні, бажано великі | Краща керованість F0/енергією | Середня |
| CycleGAN-/StarGAN-VC | Сирі/мел ознаки із цикловими втратами | Непаралельні, без пар | Частково, через додаткові голови | Середня |
| HuBERT/PPG-керовані | SSL-контент інваріантний до спікера | Непаралельні, мало прикладів | Добра з F0-ремапінгом | Висока з HiFi-GAN |

Оцінка якості системи синтезу мовлення виходить далеко за межі простого суб’єктивного опитування користувачів типу «чи звучить це природно» (Mean Opinion Score, або MOS). Хоча сприйняття природності є важливим показником, воно не охоплює всіх аспектів роботи моделі. Для комплексного аналізу необхідно враховувати також об’єктивні технічні метрики, які дозволяють кількісно оцінити різні сторони згенерованого голосу.

По-перше, оцінюють подібність синтезованого голосу до оригінального диктора. Це робиться за допомогою верифікаційних балів або метрик на основі speaker embedding – векторних представлень голосів.

По-друге, аналізують точність самого звукового сигналу, порівнюючи його спектральні характеристики з референтним записом. Для цього найчастіше використовують показник Mel-Cepstral Distortion (MCD), який вимірює спектральну відстань між двома аудіо-сигналами.

По-третє, важливо перевірити, наскільки розбірливими є слова у синтезованій мові. Це визначають через метрику Word Error Rate (WER) – частку неправильних або нерозпізнаних слів під час автоматичного розпізнавання мовлення (ASR).

На рисунку 1.2 представлено порівняння профілів підходів Voice Conversion за критеріями якості, швидкості інференсу, стабільності/узгодженості, керованості просодикою, придатності до непаралельних даних та Zero/One-shot здатності.

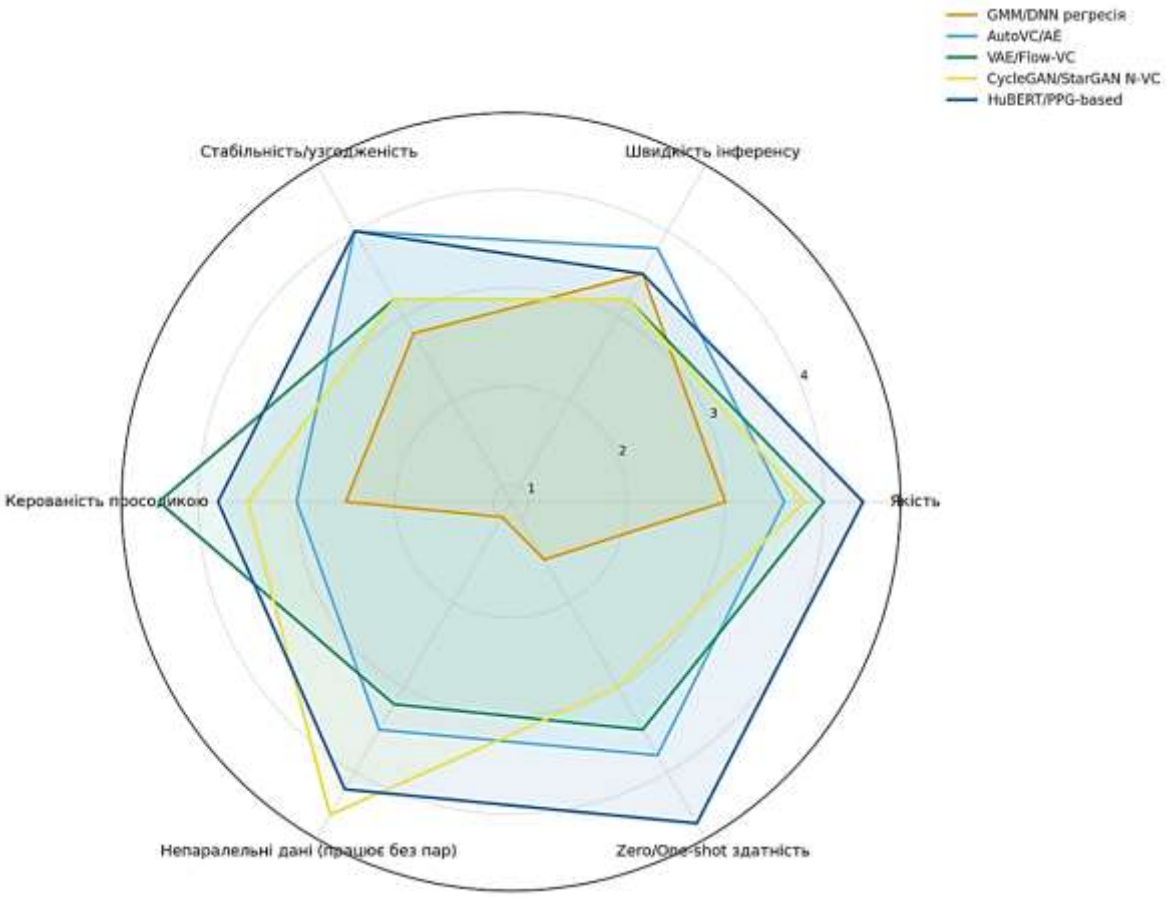


Рисунок 1.2 – Порівняльний профіль підходів Voice Conversion

Поєднання цих трьох типів показників – верифікаційних, спектральних та розбірливості – дає повну картину якості системи синтезу мовлення.

Практична побудова системи синтезу вимагає розділення змісту та унікального «цифрового зліпка» голосу (спікер-ембеддингу). При цьому, дуже важливою є сувора уніфікація всіх технічних параметрів звуку (частоти,

налаштувань спектрограм) та використання стабільного вокодера, наприклад HiFi-GAN, а для точного керування емоційними відтінками необхідно впроваджувати або приховані стилістичні латенти, або прямий контроль висоти тону й тривалості. Нарешті, для гарантованої відтворюваності результатів слід назавжди зафіксувати версії використовуваних допоміжних моделей (ASR чи HuBERT), оскільки будь-яка заміна цього «фундаменту» змінить характер помилок у даних і може порушити роботу всього конвеєра.

Таким чином, найбільш практичним та доцільним рішенням для побудови системи зміни голосу (VC) є використання контентних кодів, які не залежать від мовця, у поєднанні з окремим кодом для нового голосу та стабільним GAN-вокодером.

## Висновки до розділу 1

У цьому розділі здійснено теоретичний аналіз методів синтезу та перетворення мовлення, розглянуто еволюцію технологій від класичних до сучасних нейромережевих рішень, а також визначено вимоги до адаптації цих систем для української мови.

Встановлено, що перехід від конкатенативних та формантних методів до нейромережевих (end-to-end) архітектур дозволив подолати обмеження щодо природності звучання та гнучкості стилізації. Нейромережеві моделі забезпечують спільне навчання фонетичного контенту, просодики та тембру без необхідності ручного формулювання правил.

Порівняльний аналіз моделей Tacotron, FastSpeech та VITS показав переваги неавторегресивних підходів. Зокрема, моделі з явним прогнозуванням тривалостей (типу FastSpeech) забезпечують кращий контроль темпу та інтонації, а також вищу швидкість інференсу. Інтегровані рішення (VITS) демонструють найвищу якість генерації завдяки спільній оптимізації акустичної моделі та вокодера.

Визначено, що GAN-вокодери (зокрема HiFi-GAN) є оптимальним вибором для сучасних систем, оскільки вони забезпечують якість, порівнянну з авторегресивними моделями (WaveNet), при значно меншій латентності, що дозволяє роботу в режимі реального часу.

Обґрунтовано доцільність підходу, що базується на розділенні контентних (інваріантних) та спікер-залежних представлень. Такий принцип забезпечує більш гнучке моделювання мовлення, оскільки контентна частина відображає лінгвістичну інформацію незалежно від конкретного диктора тоді як спікер-залежна складова описує індивідуальні характеристики тембру.

Виявлено критичну важливість урахування специфіки української фонетики (йотовані голосні, позиційне пом’якшення, рухомий наголос) на етапі препроцесингу для забезпечення природної просодики.

Отже, сучасна методологія побудови систем синтезу мовлення має базуватися на end-to-end архітектурах із контрольованими просодичними параметрами та чітко уніфікованим аудіостеком (єдина частота дискретизації, узгоджені алгоритми оцінки $F_0$ та енергії). Сформовані теоретичні засади та вимоги до корпусу даних слугують основою для подальшого проєктування, навчання та експериментальної оцінки системи в наступних розділах.

# РОЗДІЛ 2
# ПРОЄКТУВАННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ СИСТЕМИ СИНТЕЗУ МОВЛЕННЯ

## 2.1 Обґрунтування вибору архітектури нейронної мережі

Архітектура системи синтезу мовлення базується на інтеграції двох взаємодоповнювальних підсистем: end-to-end моделі синтезу мовлення (TTS) для генерації аудіосигналу з тексту та конвеєра перетворення голосу (VC) для адаптації тембру. Такий гібридний підхід поєднує високу природність звучання та гнучке керування просодикою, властиві сучасним TTS-системам, із можливістю персоналізації голосу, що не вимагає створення ресурсномістких паралельних корпусів даних

Для TTS доцільно застосувати VITS-подібну архітектуру з монотонним вирівнюванням і спільною оптимізацією акустичної моделі з вокодером. Фронтенд включає нормалізацію й фонемізацію тексту, далі текстовий енкодер проєктує послідовність у латентний простір; нормалізуючі флови формують керований латент варіативності, а генератор у поєднанні з дискримінаторами (часовими/частотними) навчається відтворювати сигнал із високою природністю. Така схема зменшує залежність від зовнішніх вирівнювачів тривалості та уніфікує цільову функцію для акустики й хвилі [14].

Підсистема VC використовує інваріантний до мовця контентний представник (на кшталт HuBERT/BN-ознаки) разом із окремим спікер-ембеддингом і явним моделюванням висоти тону $F_0$. Генератор виконує умовлене перетворення ознак джерела у простір цільового мовця, після чого стабільний GAN-вокодер відновлює часовий сигнал. У межах спільного продуктового рішення TTS і VC поділяють однаковий аудіо-препроцесинг та параметри ознак, що спрощує підтримку, прискорює інференс і знижує кількість артефактів на довгих послідовностях для правильної роботи підсистеми.

## 2.2 Формування та препроцесинг навчального корпусу для моделі синтезу мовлення

Основною метою підготовки даних є формування чистого, збалансованого та відтворюваного корпусу, придатного для навчання як end-to-end TTS, так і систем перетворення голосу (VC). Процес передбачає сувору уніфікацію форматів, нормалізацію транскриптів та фіксацію зв'язків «аудіо ↔ текст/метадані» у маніфестах зі стабільною структурою. Розподіл на вибірки (train/val/test) здійснюється заздалегідь із контролем перекриття мовців для коректної оцінки узагальнюваності моделі [15].

Аудіозаписи приводяться до уніфікованого формату: моно, частота дискретизації 24 кГц, розрядність 16-біт PCM. Виконується нормалізація гучності (орієнтир – −23 LUFS), видалення надмірних пауз і слабких шумів, а також обрізка (тримінг) фраз до робочої довжини. Етап очищення включає дедуплікацію та видалення проблємних файлів (кліпінг, реверберація, сторонні звуки) ще до екстракції ознак. Щоб уникнути зміщень у стилі чи темпі, застосовується стратифікація даних за мовцями та тривалістю записів. Усі параметри препроцесингу документуються в єдиній конфігурації [16].

Текстові транскрипти проходять процедуру нормалізації, що охоплює розгортання чисел, одиниць виміру, абревіатур та обробку власних назв; за потреби виконується фонемізація згідно з орфоепічними нормами. Фінальні маніфести мають уніфікований формат, сумісний із завантажувачами даних, і містять шляхи до файлів, текст (або фонеми), ідентифікатор мовця, тривалість та належність до конкретного спліта (вибірки). Це гарантує прозорість перевірки якості та стабільну відтворюваність експериментів.

Для забезпечення сумісності всіх підсистем (від навчання акустичної моделі до інференсу VC) використовується єдина незмінна конфігурація аудіопроцесора. Вона визначає: параметри STFT – розмір перетворення (fft_size), крок (hop_length), довжину вікна (win_length) та тип віконної функції; мел-спектрограми – кількість смуг (n_mels), частотний діапазон

(fmin/fmax) та параметри нормування; просодичні ознаки – метод вилучення основного тону (F0) та міток озвученості (voicing).

Сталість цих параметрів дозволяє уникнути систематичних зсувів та зменшити кількість артефактів. Деталі ініціалізації аудіопроцесора та параметри екстракції наведено на рисунку 2.1.

```
> Using model: vits
> Setting up Audio Processor...
| > sample_rate:24000
| > resample:False
| > num_mels:80
| > log_func:np.log10
| > min_level_db:0
| > frame_shift_ms:None
| > frame_length_ms:None
| > ref_level_db:None
| > fft_size:1024
| > power:None
| > preemphasis:0.0
| > griffin_lim_iters:None
| > signal_norm:None
| > symmetric_norm:None
| > mel_fmin:0
| > mel_fmax:None
| > pitch_fmin:None
| > pitch_fmax:None
| > spec_gain:20.0
| > stft_pad_mode:reflect
| > max_norm:1.0
| > clip_norm:True
| > do_trim_silence:False
| > trim_db:60
| > do_sound_norm:False
| > do_amp_to_db_linear:True
| > do_amp_to_db_mel:True
| > do_rms_norm:False
| > db_level:None
| > stats_path:None
| > base:10
| > hop_length:256
| > win_length:1024
```

Рисунок 2.1 – Ініціалізація аудіопроцесора та параметри екстракції ознак

Логічним завершенням етапу підготовки даних є програмна реалізація механізмів їх зчитування та пакетування. Для забезпечення коректної подачі нормалізованих аудіозаписів і транскриптів у нейромережу використовується спеціалізований завантажувач, налаштування якого мають чітко узгоджуватися

зі структурою вхідних файлів метаданих. Фрагмент коду, що демонструє формат полів маніфесту та відповідну конфігурацію класу завантажувача даних, наведено на рисунку 2.2.

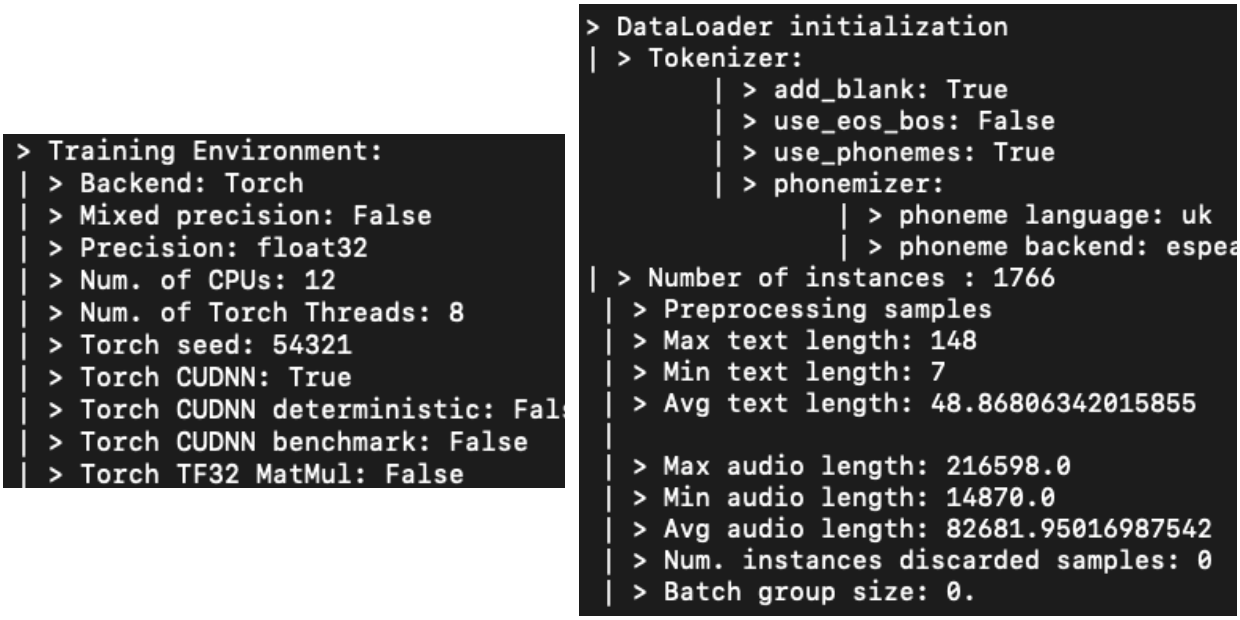


Рисунок 2.2 – Структура маніфестів і конфігурація завантажувача даних

На завершення підготовчої стадії, навчальний корпус перевіряється на відповідність цільовим метрикам якості (чистота, покриття фонетики, баланс мовців) і лише після цього фіксується як «заморожена» версія для навчання, валідації та подальших порівнянь між моделями.

### 2.3 Алгоритмічна реалізація процесу навчання моделі

Процедура навчання нейронної мережі для синтезу мовлення організується як точний і відтворюваний ланцюжок дій (конвеєр), де для гарантії повторюваності фіксуються стартові налаштування та використовуються однакові параметри для обробки звуку на всіх етапах. Щоб навчання було швидким і стабільним, воно виконується з високою ефективністю (у змішаній точності), але з постійним контролем: різкі, невдалі

зміни обмежуються (gradient clipping), а обчислення об’єднуються для більшої стійкості; також періодично зберігаються версії моделі, їх параметри усереднюються (EMA), що робить фінальну якість голосу більш рівномірною між усіма збереженими етапами. Нарешті, перевірка якості та запис результатів проводяться через строго однакові проміжки часу, що дозволяє коректно порівнювати динаміку навчання в різних робочих сесіях.

Дані надходять у модель через стратифікований завантажувач із фіксованими розмірами кадрування і стабільним співвідношенням train/val/test; для довгих фраз застосовується обрізання або розбиття їх на блоки фіксованої тривалості, що утримує GPU-пам’ять у межах цільових лімітів. Швидкість збіжності забезпечується комбінацією оптимізатора AdamW і косинусного планувальника з розігрівом, а регуляризації на рівні акустичної моделі та вокодера поєднуються зі збереженням «чистої» мел-репрезентації (без зміни параметрів STFT протягом усіх експериментів) [17].

Моніторинг якості реалізовано через TensorBoard. Відслідковуються загальні й компонентні втрати, навчальні/візуальні приклади лог-мелів із підписами k-step, а також хвильові форми «реальне/згенероване/різниця». Для полегшення трасування використовуються контрольні піднабори прикладів з фіксованими текстами й мовцями, що дозволяє зіставляти зближення формантної структури та контуру $F_0$ під час навчання. Для інтерпретації динаміки інколи робляться знімки проміжних станів на реперних кроках, зокрема у фазі стабілізації уваги та вокодера.

Візуалізація панелі моніторингу відображає стартові параметри процесу навчання, включно з початковими значеннями гіперпараметрів, конфігурацією оптимізатора та розміром батчу. На графіках фіксується динаміка зміни швидкості навчання упродовж перших епох, що дозволяє контролювати адаптацію моделі до даних. Додатково подається поведінка окремих складових цільової функції на початкових ітераціях – наприклад, варіаційного розподілу та регуляризаційних термів, що забезпечує оперативне виявлення нестабільності, наведена на рисунку 2.3.

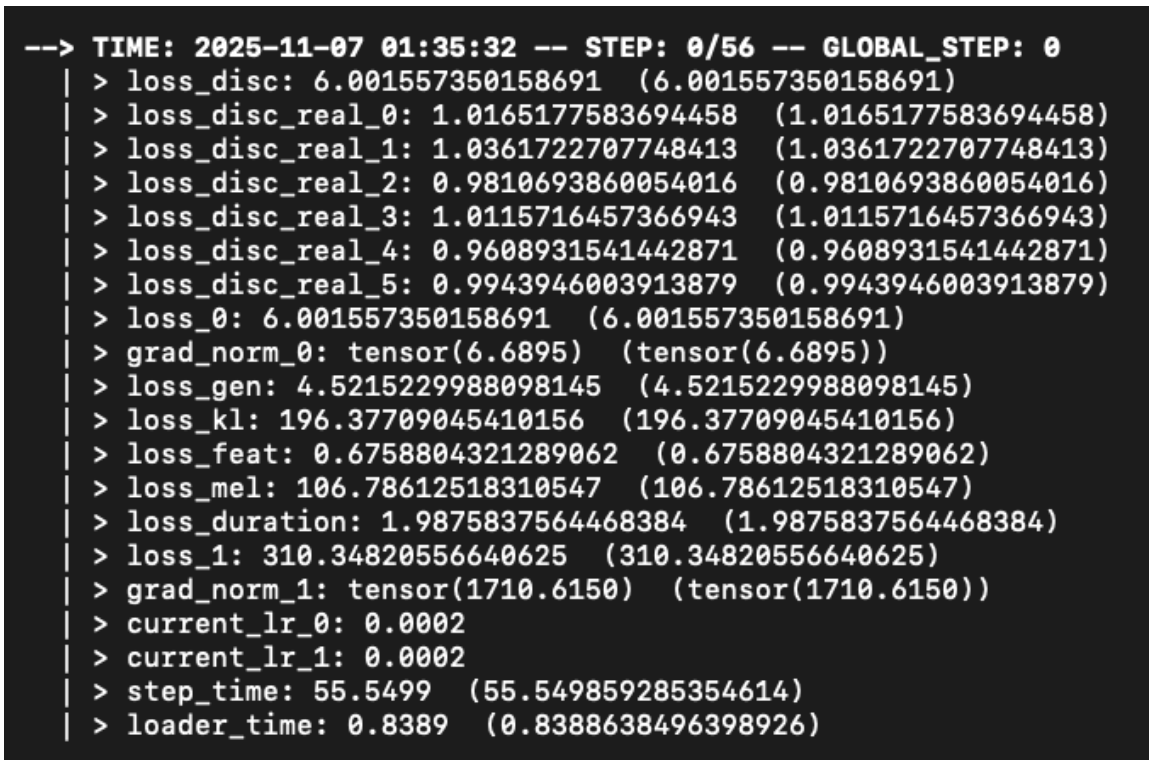


Рисунок 2.3 – Запуск навчання нейронної мережі для синтезу мовлення

Для ілюстрації проміжного стану моделі, де вже простежується формування акустичних ознак, але ще присутні характерні шуми розмиття, наведено знімок метрик та спектрограм на етапі 25-го кроку (рисунок 2.4)

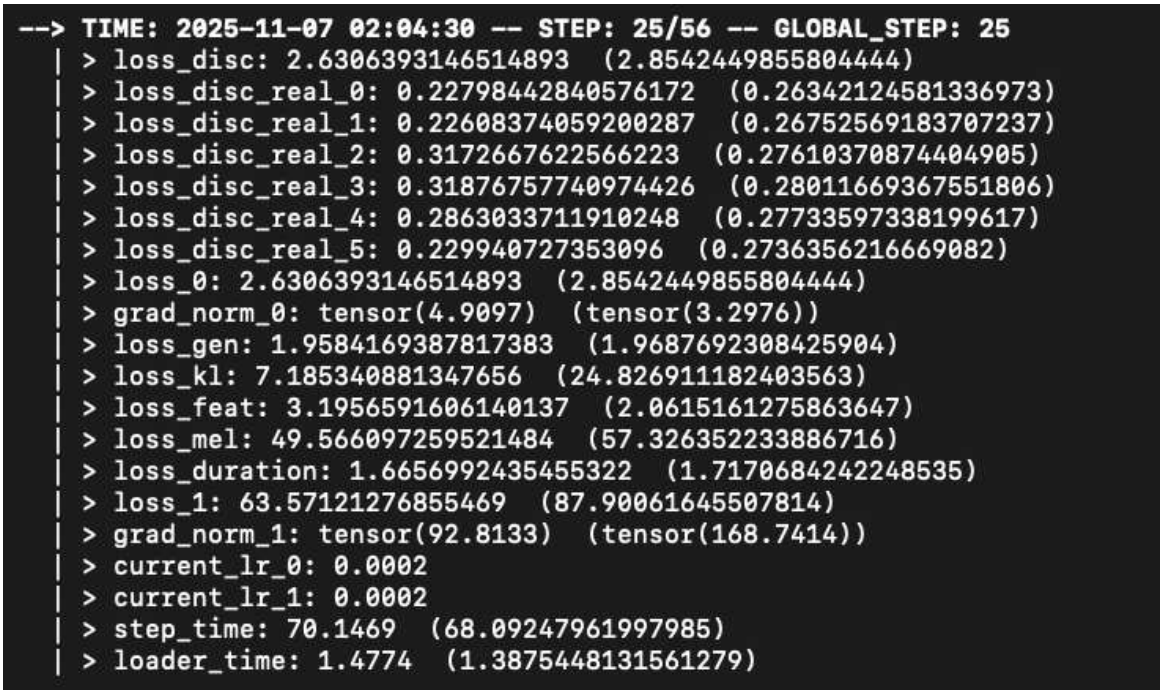


Рисунок 2.4 – Динаміка збіжності моделі на проміжному етапі (крок 25/56)

Подальший прогрес оптимізації, що характеризується суттєвим зменшенням спектральних артефактів та уточненням деталей генерації на пізніх ітераціях, зафіксовано на рисунку 2.5

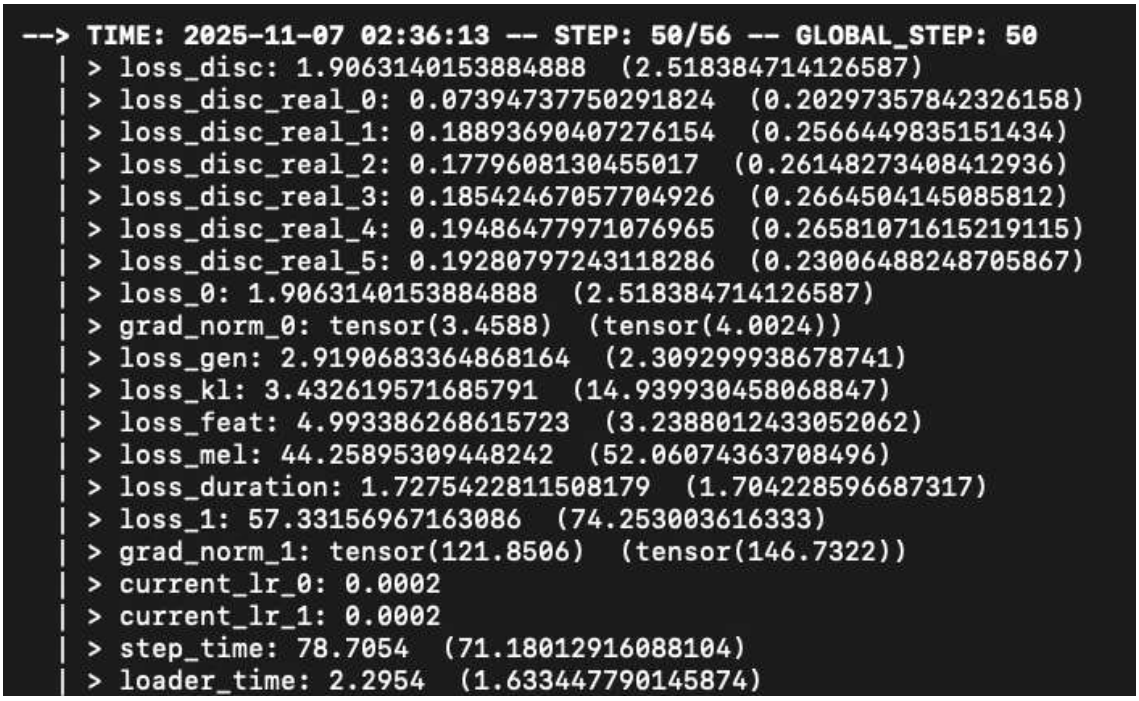


Рисунок 2.5 – Динаміка збіжності на пізньому етапі (крок 50/56): стабілізація спектральних артефактів

Візуалізація мел-спектрограм у TensorBoard є важливим інструментом для якісного аналізу систем синтезу мовлення. За допомогою цього інструментарію можна одночасно відобразити згенеровані мел-спектрограми та цільові еталонні спектрограми, що дозволяє візуально порівняти текстуру, форму та динаміку звукового сигналу. Для глибшого розуміння відмінностей між синтезованим та оригінальним звуком TensorBoard надає можливість побудови карти різниці (помилки), яка відображає спектральні відхилення у вигляді теплової карти, що підкреслює ділянки найбільших неточностей

Це дозволяє розробникам швидко виявляти проблемні місця в синтезі, такі як втрати деталей у формантах чи неправильне відтворення експресії, та оптимізувати модель відповідно, щоб вона могла працювати без нарікань, показано на рисунку 2.6.

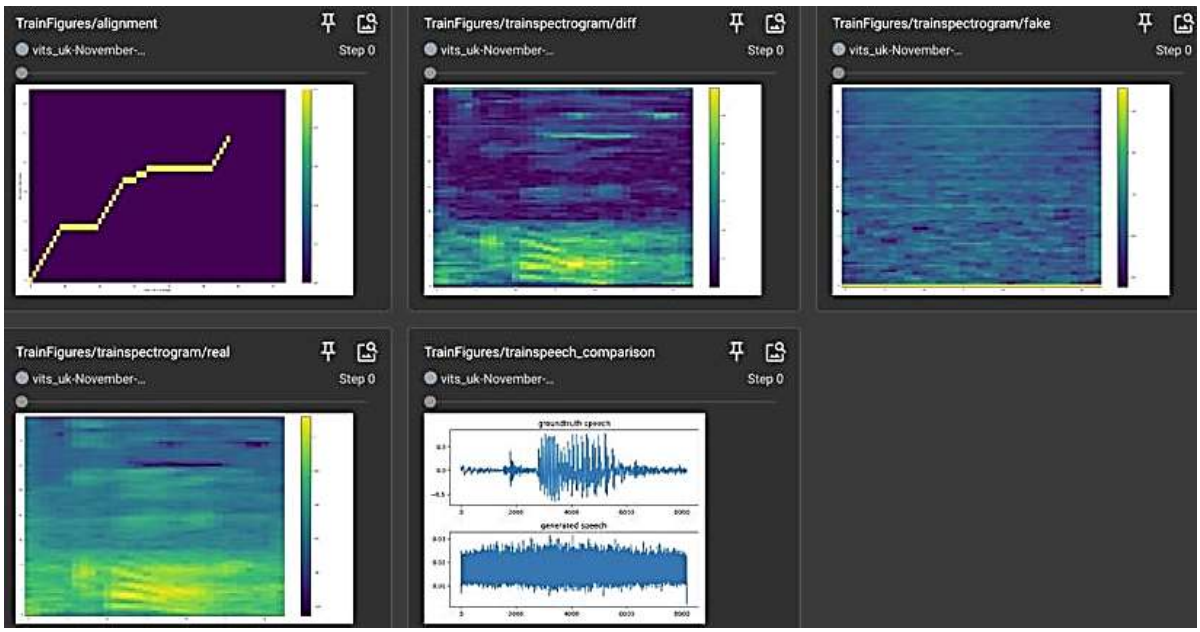


Рисунок 2.6 – TensorBoard / Images: порівняння «реальне – згенероване – різниця» на контрольних фразах

Оцінювання під час навчання відокремлене від фінальної експертизи, але покриває ключові орієнтири: спектральні розбіжності (MCD), точність реконструкції висоти тону (F0-RMSE), а також інтегральні індикатори зрозумілості через WER/CER стороннього ASR. Валідаційний завантажувач формує відтворювану послідовність прикладів, а підсумковий борт метрик візуалізує прогрес по кожному чекпоінту. Додатково фіксується реальний часовий фактор інференсу (RTF) для вибраного апаратного профілю, що дозволяє відкидати моделі, непридатні для оперативних сценаріїв.

Схема валідації для систем синтезу мовлення будується на формуванні відтворюваних пакетів даних – фіксованих наборів аудіо-текст пар з заданими seed'ами для рандомізації та time-sorted split (80/20 по користувачах), що гарантує однакові умови тестування на кожній епосі навчання. Паралельно створюються списки контрольних прикладів (benchmark sentences) – репрезентативні фрази для перевірки ключових аспектів якості: просодії, тембру, розбірливості та стабільності голосу. У процесі навчання ці приклади досить часто використовуються для відстеження та для спостереження за змінами у просодичних характеристиках та відтворюваності тембру (рисунок 2.7).

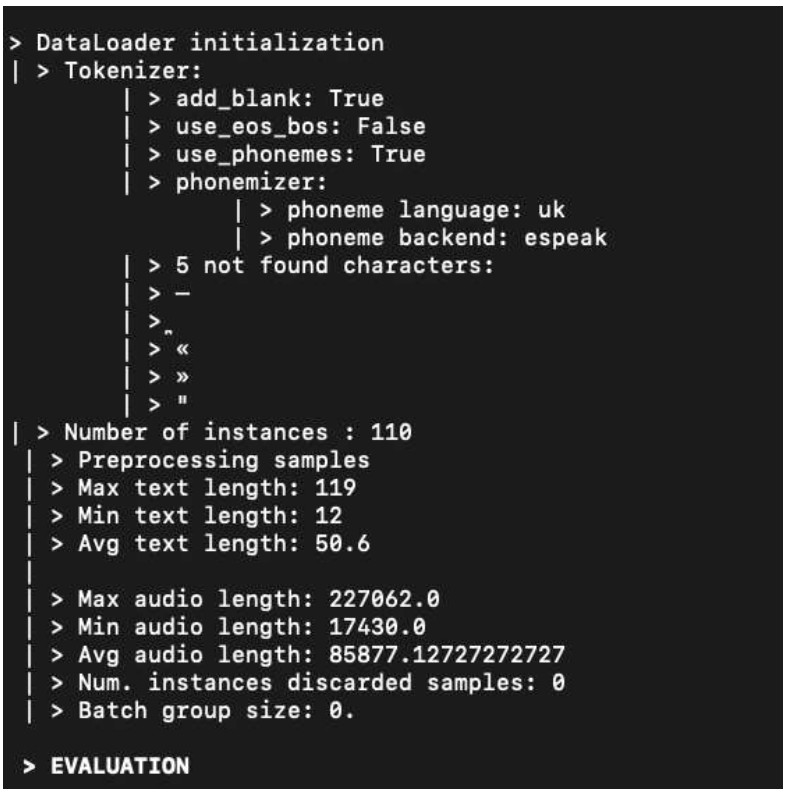


Рисунок 2.7 – Валідаційний конвеєр: формування батчів і відтворюваний список контрольних прикладів

Зведені значення об’єктивних метрик якості (спектральної відстані, похибки висоти тону та рівня розпізнавання тексту), що дозволяють оцінити загальну ефективність навчання моделі у динаміці, представлено на рисунку 2.8.

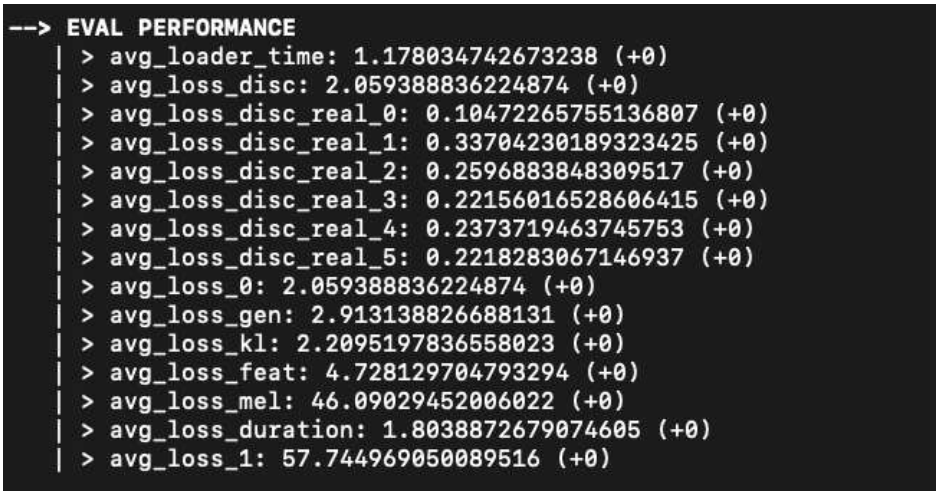


Рисунок 2.8 – Зведені метрики валідації: MCD, F0-RMSE, WER/CER за проміжними версіями моделі

Деталізація роботи механізму уваги, що демонструє коректність часового узгодження тексту та аудіо, разом із відповідними хвильовими формами та енергетичними картами наведена на рисунку 2.9.

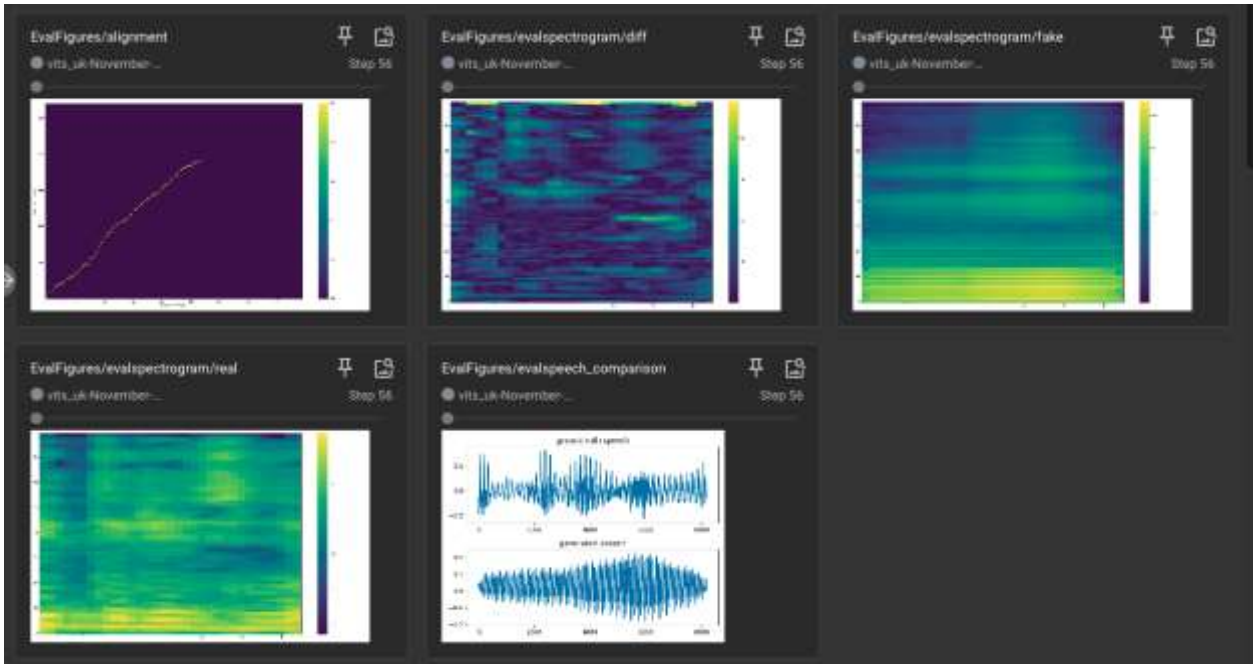


Рисунок 2.9 – TensorBoard / Alignment & Waveplots: вирівнювання, хвильові форми та енергетичні карти

Як видно з наведених далі даних, зафіксований набір параметрів забезпечує компроміс між високою якістю генерації (завдяки комбінації змагальних втрат та багатошкальних дискримінаторів) та ресурсоефективністю (використання змішаної точності FP16 та цільовий RTF ≤ 0,1). Застосування механізмів стабілізації, таких як експоненційне ковзне середнє (EMA) та обмеження градієнтів, дозволяє мінімізувати ризики розбіжності процесу навчання, створюючи надійну технічну базу для подальшого аналізу ефективності моделі.

Для забезпечення повної відтворюваності експериментальних досліджень та уніфікації технічних умов функціонування системи, ключові гіперпараметри навчання та інференсу було зведено в єдиний реєстр. Обрані конфігурації базуються на результатах попередніх оптимізаційних тестів і є

спільними для обох підсистем (TTS та VC), що гарантує сумісність акустичних ознак та стабільність обчислювального процесу. Детальний перелік базових налаштувань, що визначають архітектуру експерименту, представлено в таблиці 2.1.

Таблиця 2.1 – Базові налаштування навчання та інференсу

| Параметр | Значення | Примітка |
|---|---|---|
| Частота дискретизації (SR) | 24000 Гц | Єдина для TTS і VC |
| Мел-банки / STFT | n_mels = 80; fft_size = 1024; hop = 256; win = 1024; fmin/fmax = 0/12000 | Незмінні в усіх експериментах |
| Оптимізатор | AdamW ($\beta_1 = 0{,}8$; $\beta_2 = 0{,}99$; wd = 1e−2) | Стійкість до перенавчання |
| Планувальник LR | Cosine decay з warmup 4000 кроків; $LR_0 = 2e{-}4$ | Швидка збіжність без піків |
| Розмір батча | train = 16; eval = 8; grad_accum = 2 | Баланс швидкості та VRAM |
| Змішана точність | FP16 (AMP) | Прискорення інференсу/навчання |
| Обмеження градієнта | clip = 1,0 | Запобігання вибуху градієнтів |
| EMA параметрів | коеф. 0,999 | Стабільніші чекпоінти |
| Втрати (узагальнено) | L1/L2 на мел + Multi-STFT + Adv + Feature Matching + KL | Узгодження акустики і хвилі |
| Дискримінатори | Multi-Scale + Multi-Period | Контроль часової та тональної структури |
| Критерій зупинки | плато вал. STFT-втрат + стабілізація метрик | Уникає перенавчання |
| Цільовий RTF | ≤ 0,1 на GPU класу consumer | Придатність до інтерактивності |

Отже, процес навчання системи синтезу мовлення має бути побудований так, щоб гарантувати: по-перше, можливість точного й повного відтворення результатів експерименту; по-друге, стабільність самого процесу оптимізації, що запобігає розбіжності або деградації моделі; і по-третє, постійний контроль якості синтезованого голосу за низкою об’єктивних і суб’єктивних метрик. Такий підхід дозволяє не лише своєчасно виявляти відхилення в роботі моделі, а й поступово вдосконалювати параметри навчання. У результаті можна одночасно усунути недоліки звучання.

## 2.4 Інтеграція підсистеми переозвучення аудіофалів

Механізм переозвучення аудіофайлів голосом конкретної особи реалізується як послідовність етапів: від виділення контентних ознак із сигналу джерела до реконструкції хвилі з тембральними характеристиками цільового мовця. Усі обчислення виконуються на єдиній частоті дискретизації 24 кГц, із тією самою конфігурацією ознак, що застосована в підсистемі TTS, щоб уникнути зсувів між треками. Синхронізована конфігурація забезпечує однакові параметри STFT/мел-представлення та метод оцінки $F_0$, завдяки чому конверсія стабільно відтворює інтонаційні контури без появи «зіпсованих» пауз або дзижчання у високочастотному діапазоні.

На вхід подається мовлення «джерела», з якого екстрагуються контентні, інваріантні до мовця представлення (на кшталт BN/HuBERT), а також ознаки висоти тону й енергії; паралельно з репозитарію або каталогу ембеддингів завантажується спікер-вектор «цілі». Далі умовлений перетворювач картує контент у простір цільового мовця, коригуючи контур $F_0$ через нормалізацію в лог-просторі та ремапінг під статистики «цілі», після чого стабільний GAN-вокодер (HiFi-GAN-клас) відновлює часовий сигнал. Такий розподіл на «контент» і «спікер» дає змогу працювати в one-/zero-shot режимах і не вимагає паралельних пар «джерело–ціль» у навчанні [18].

У реалізації особливу увагу було приділено стабільності керування та управління інтонацією. Оскільки саме $F_0$ відповідає за сприйняття висоти та «мелодики» висловлювання, його контур не перезаписується довільно, а нормалізується відносно статистик джерела та переноситься у статистики цілі, що дозволяє зберегти ритміку, не створюючи ефекту «гелієвого» голосу. Для монотонних і «скандованих» фрагментів увімкнена відсічка неозвучених кадрів (voicing), аби вокодер не «дозвучував» шум у місцях пауз; для дуже насичених приголосних використовується м’який high-shelf, який приглушує потенційний металевий «дзвін» без втрати чіткості дикції у озвученому тексті (рисунок 2.9).

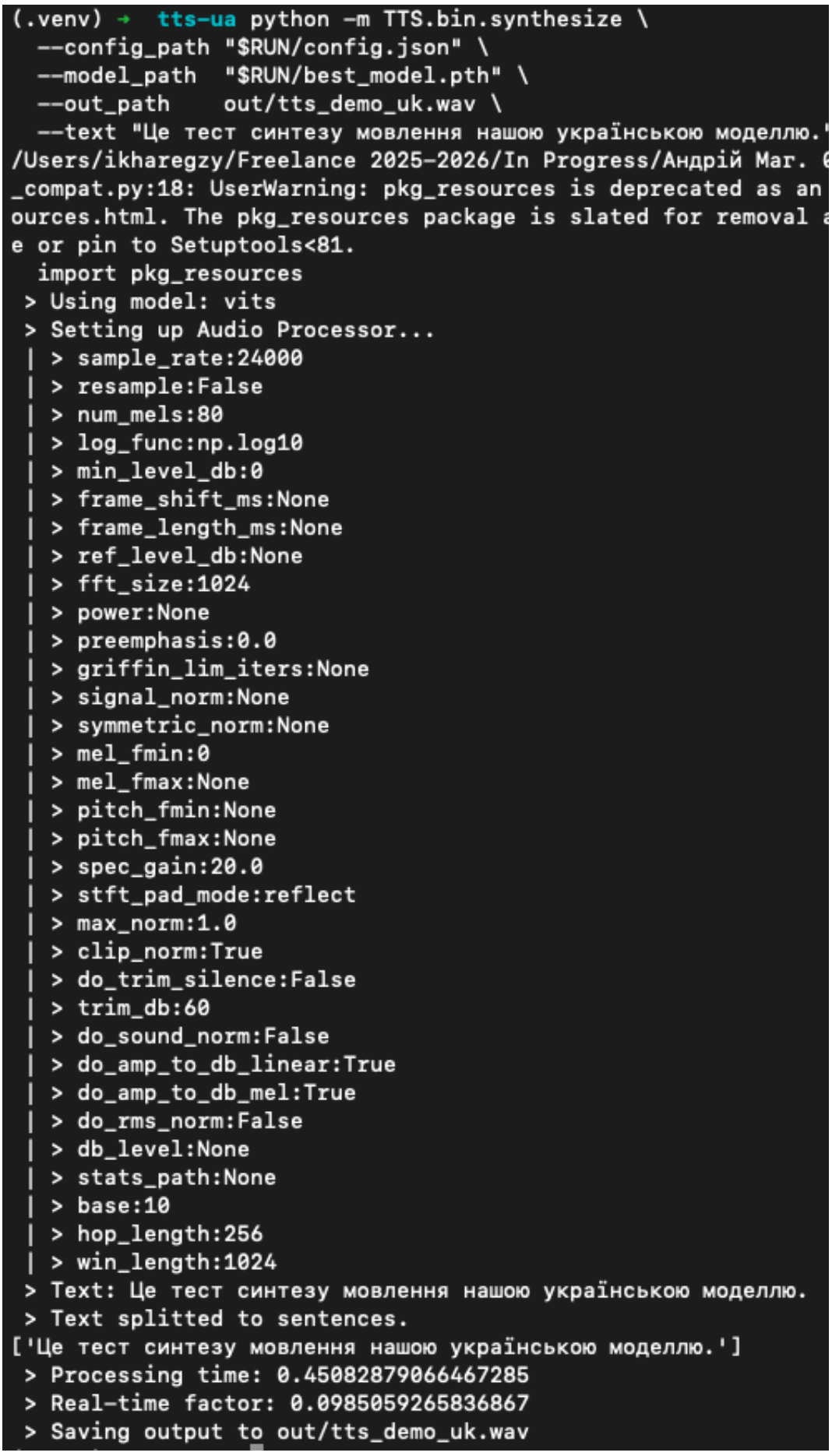


Рисунок 2.9 – CLI-запуск інференсу (TTS): параметри аудіопроцесора та RTF

Операційно переозвучення доступне через окрему CLI-команду з явним зазначенням шляху до аудіо-джерела, ідентифікатора/вектора цільового мовця, каталогу виводу й конфігурації ознак; у журнали автоматично заносяться тривалість треку, час обробки й обчислений real-time factor (RTF). Значення RTF < 0,1 для типового GPU-класу «consumer» інтерпретується як «достатньо для інтерактивних сценаріїв», а стабільність вимірюється повторними прогонами на «еталонному» контрольному наборі фраз з різними темпами й регістрами (рисунок 2.10).

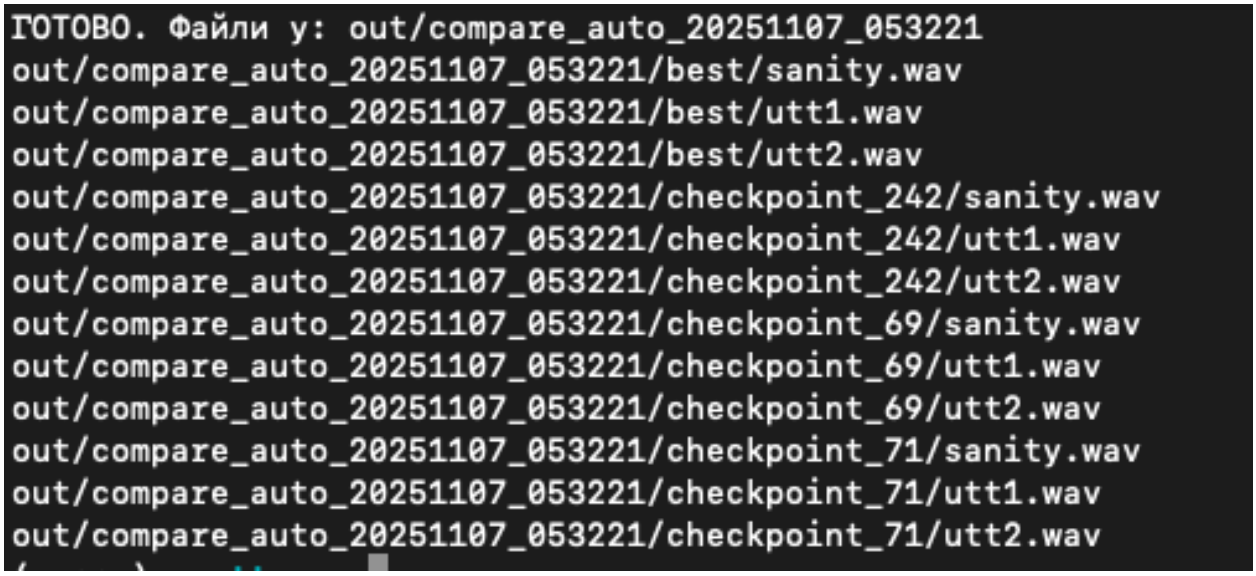


Рисунок 2.10 – Структура контрольних виходів: файли «converted.wav», журнали сесії та діагностичні спектрограми

Для забезпечення відтворюваності та подальшої перевірки результати структуровано у каталозі виводу: поруч із «converted.wav» для кожного прикладу зберігаються лог-файли, версії конфігурацій і, за потреби, діагностичні лог-мел-спектрограми. Така організація дає змогу пов’язати артефакт із конкретною версією налаштувань і оперативно повторити інференс зі зміненими параметрами. Усі етапи – від нормалізації входів до збереження виходів – використовують однакові параметри ознак, що унеможливлює конфігурації між підсистемами та знімає питання про узгодженість із навчальним контуром.

## 2.5 Оптимізація моделі для підвищення якості мовлення

Висока якість генерації мовлення досягається завдяки точному налаштуванню правил навчання, які змушують модель одночасно звертати увагу на важливі деталі та створювати результати, що сприймаються як справжні. Щоб забезпечити моделі широкий простір для творчості та запобігти зацикленню на одному рішенні, використовується додатковий механізм м’якого контролю ідей. Процес навчання робиться стійким і плавним двома способами: різкі, невдалі кроки навчання завжди обмежуються, а сама модель постійно згладжується усередненням. Це гарантує, що фінальний результат буде чистим без жодного шуму чи зернистості, а якість залишатиметься високою на всіх етапах. Нарешті, швидкість навчання контролюється спеціальною програмою, яка починає повільно, а потім поступово знижує інтенсивність, забезпечуючи точне й плавне знаходження найкращого рішення.

Керованість просодикою забезпечується архітектурним рішенням, у якому за прогнозування частоти основного тону ($F_0$) та енергії відповідають відокремлені обчислювальні модулі (незалежні предиктори). Щоб уникнути ефекту неприродної ритмічної однорідності («скандування»), під час навчання до тривалостей фонем додаються випадкові варіації, а згладжування пауз дозволяє стабілізувати темп у довгих фразах.

У задачах перетворення голосу (VC) контур $F0$ нормалізується в логарифмічному масштабі та адаптується до діапазону висоти цільового мовця шляхом статистичного перетворення. Це дозволяє зберегти оригінальну інтонацію, уникнувши неприродних спотворень тембру (ефекту надто високого чи низького голосу). Крім того, використання ідентичних налаштувань частотно-часового аналізу (STFT) у всіх компонентах системи гарантує сумісність сигналів та усуває систематичні помилки обробки.

Регуляризації залишаються помірними: невеликий dropout у текстовому енкодері, легкий джиттер амплітуди на мел-послідовностях без зміни STFT-

сітки, відмова від агресивних RIR-аугментацій, що погіршують MOS на чистих доріжках. Для вокодера використовується багатошкальна STFT-втрата із щільнішим контролем високих частот і feature matching; мультидискримінатори за масштабами й періодами дисциплінують як часову, так і тональну структуру сигналу. На етапі інференсу зниження RTF досягається змішаною точністю, експортом у TorchScript/ONNX, об'єднанням дрібних операцій та сегментацією довгих фрагментів зі згладженими швами; вимірювання реального часу прив'язується до сценарію CLI, подібного до показаного на рисунку 2.9 [19].

Контроль якості включає порівняння триптихів лог-мелів «джерело → ціль → конвертований», перевірку верифікаційних балів схожості, стабільний ASR-декодер для WER/CER і повторювані MOS-сесії з однаковими слухачами. Якщо локальне покращення однієї метрики супроводжується погіршенням зрозумілості чи подібності, зміни відкочуються; пріоритет надається інтегральній якості мовлення та відтворюваності результатів на контрольному піднаборі. Сукупність цих заходів забезпечує більш чисту спектральну картину, природнішу просодику й швидкодію інференсу без компромісів щодо стабільності.

Стимули формуються трійками для кожного тексту: природний еталон, синтез TTS та переозвучення VC. Усі файли моно, 24 кГц, 16-біт PCM із нормуванням до −23 LUFS; тиші уніфіковано, записи з кліпами або вираженими артефактами вилучено. Тривалість фраз становить 3–7 с, що забезпечує достатнє фонетичне покриття й різні просодичні патерни. Для подальших обчислень застосовується однакова конфігурація STFT/mel-ознак, ідентична навчальній, аби уникнути систематичних зсувів між етапами.

Об'єктивні метрики розраховуються на тих самих парах «еталон ↔ система», що подаються учасникам у суб'єктивних тестах. MCD обчислюється на наборах MFCC, отриманих зі спільної mel-репрезентації (D=24, без $c_0$); F0-RMSE оцінюється у центах на озвучених кадрах за лог-шкалою висоти; WER/CER – після уніфікованої текстової нормалізації на виході одного ASR-

декодера [20]. Агрегація здійснюється на «замороженому» контрольному наборі з фіксацією версій конфігурацій та параметрів ознак.

## Висновки до розділу 2

Розроблено комплекс програмних рішень для синтезу та повторного озвучення мовлення, який поєднує end-to-end TTS на базі VITS із підсистемою voice conversion, спроєктованою навколо контентно-інваріантних ознак і спікер-ембеддингів. Єдиний аудіостек із узгодженими параметрами ознак (частота дискретизації, STFT/mel, метод оцінки висоти тону) застосовано для обох підсистем, що усунуло систематичні зсуви між модулями і знизило ризик появи артефактів на стику «акустика-вокодер». Архітектурне рішення забезпечує керованість просодикою через явні представлення $F_0$, енергії і підтримує one/zero-shot режим переозвучення без вимоги до паралельних пар «джерело-ціль».

Підготовлено відтворюваний корпус із фіксованими маніфестами, нормалізацією рівня, очищенням від шумових артефактів та стратифікованими сплітами. Це дозволило стабілізувати навчання і виключити витік даних між етапами. Конвеєр тренування реалізовано зі змішаною точністю, AdamW та планувальником із розігрівом; до цільової функції інтегровано багатошкальні STFT-втрати, адверсаріальні складові та feature matching у поєднанні з регуляризацією латентного простору. Застосовано стабілізуючі техніки на кшталт gradient clipping, EMA та усереднення останніх чекпоінтів, що зменшило коливання якості між близькими версіями моделі.

Розроблено операційний пайплайн інференсу з уніфікованими CLI-командами, журналюванням параметрів сесій і структурованим каталогом вихідних матеріалів. Для довгих фрагментів реалізовано сегментацію зі згладженими швами, що мінімізує клацання й фазові невідповідності; для macOS і GPU-сценаріїв передбачені оптимізації на рівні обчислювальних

бекендів та експорту моделей, що забезпечує режим роботи, близький до реального часу. Єдина конфігурація препроцесингу для джерела та цілі в підсистемі переозвучення усунула типові «кліки» й вузькосмуговий дзвін, а лог-нормалізація та ремапінг $F_0$ під статистики цільового мовця підвищили природність інтонацій.

Оптимізаційні рішення систематизовано у вигляді практичних правил: обережне балансування перцептивних та адверсаріальних втрат, помірні аугментації без агресивних реверберацій, дисципліна даних і незмінність аудіоконфігурації на всіх етапах. У результаті отримано стабільний і відтворюваний технологічний ланцюг, який гнучко масштабують під різні корпуси та сценарії застосування і який готовий до формальної оцінки якості в експериментальному розділі.

# РОЗДІЛ 3
# ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ СИСТЕМИ

## 3.1 Організація експерименту та методологія тестування

Експеримент з тестування синтезованого мовлення побудовано за схемою, де кожен учасник оцінює всі варіанти стимулів. Таким чином, зменшуються відмінності між оцінками різних учасників і підвищується здатність виявляти різницю між протестованими варіантами озвучення. Порівнюються три варіанти озвучення: природний еталон, синтез TTS (VITS-підсистема) та переозвучення VC для того самого тексту. Зразки для оцінювання подані у вигляді коротких фраз тривалістю 3–7 секунд, підібраних так, щоб охопити типові фонетичні послідовності української мови та різні інтонаційно-ритмічні схеми. Для зменшення ефекту звикання до матеріалу застосовано протокол рандомізації порядку відтворення та схему латинського квадрата, побудовану на рівні наборів фраз.

У дослідженні взяли участь 24 дорослі респонденти віком 18–45 років без заявлених порушень слуху; обов’язкова умова – досвід щоденного користування навушниками або акустичними системами та базова цифрова грамотність для виконання інструкцій прослуховування. Рекрутинг здійснювався серед студентів і співробітників ВНЗ, участь добровільна, безоплатна; отримано інформовану згоду, персональні дані не збирались. Для зменшення зсувів у сприйнятті просили уникати участі осіб, які професійно працюють із аудіовиробництвом; статево-вікова структура максимально збалансована. Критерії включення/виключення та узгоджена інструкція зафіксовані в протоколі експерименту [21].

Підготовка стимулів виконувалася в єдиному аудіостеку: усі записи моно з частотою дискретизації 24 кГц та глибиною 16 біт; рівні нормовано до −23 LUFS із збереженням пікового headroom. Для кожної фрази існує трійка

варіантів «еталон / TTS / VC», отриманих із одного тексту; для VC додатково перевірено правильність відповідності цільовому мовцю через верифікаційний скор у просторі спікер-ембеддингів. Набір стимулів поділено на тренувальний блок із прикладами та основний тестовий блок; тренувальний блок не входить до аналізу.

Умови прослуховування стандартизовані: тихе приміщення (рівень фонових шумів не вище 30–35 dBA), закриті динамічні навушники з відтворенням не гірше 20 Гц–20 кГц, гучність, відкалібрована на рожевому шумі до комфортного рівня без спотворень. Під час сесії заборонено змінювати гучність; між стимулами – короткі технічні паузи, щоб запобігти слуховій втомі. Для онлайн-слухачів передбачено обов'язковий пре-тест на виявлення неправильного підключення, моно/стерео та «перемикання» каналів; результати учасників, які не пройшли пре-тест, відсікаються.

Організаційно експеримент ділиться на інструктаж, тренувальні приклади, основний етап оцінювання та коротке опитування у форматі анкети про умови прослуховування (тип навушників, суб'єктивна зручність, рівень відволікань). Усі ідентифікатори стимулів і відповіді логуються; порядок подачі стимулів унікальний для кожного учасника. Використовуються об'єктивні метрики: для тих самих фраз окремо обчислюються MCD, F0-RMSE і WER/CER, що дозволяє зіставляти суб'єктивні оцінки з автоматичними показниками у підрозділах 3.2–3.3, а також відтворити експеримент у разі перевірки.

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(c_d - \widehat{c_d})^2}, \qquad (3.1)$$

де: $c_d$ – це характеристики (числа), що описують спектр реального голосу (еталон); $\widehat{c_d}$ – це ті самі характеристики числа, але які використовуються для синтезованого голосу.

Формула (3.1) – мел-кепстральне спотворення (MCD). Ця формула обчислює «відстань» між двома звуками. Тобто ми беремо різницю між ними, підносимо до квадрата і сумуємо. Чим менше число вийде в результаті, тим ближче синтезований голос до реального за звучанням (тембром).

Тут наводиться формула середньоквадратичної помилки (RMSE) для контуру частоти основного тону ($F_0$). Вона оцінює точність відтворення інтонації.

$$RMSE_{F_0} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(f_i - \widehat{f_i}\right)^2}, \tag{3.2}$$

де: $f_i$ – висота тону (нота) у реальному записі в конкретний момент часу; $f_i$ – висота тону в синтезованому записі в той самий момент.

Формула (3.2) усереднює всі відхилення по всій фразі. Чим менше отримане значення, тим точніше синтезатор скопіював інтонацію та емоційний малюнок мовця. Ця формула показує, наскільки сильно синтезатор «сфальшивив» мелодію мови.

Метрики WER (Word Error Rate – помилка на рівні слів) та CER (Character Error Rate – помилка на рівні символів) є стандартом для оцінки систем розпізнавання мовлення (ASR), але в контексті синтезу мовлення (TTS) вони використовуються для оцінки розбірливості (intelligibility).

Суть методу: синтезований звук «проганяють» через систему розпізнавання мовлення і порівнюють отриманий текст із вихідним.

Метрика WER показує відсоток слів, що були розпізнані неправильно. Вона базується на відстані Левенштейна (мінімальній кількості дій, необхідних для перетворення одного тексту в інший).

$$WER = \frac{S + D + I}{N} \cdot 100\% \tag{3.3}$$

де: $S$ – кількість слів, які були замінені на інші (наприклад, «хата» замість «тато»); $D$ – кількість слів, які були пропущені (були в еталоні, але зникли у розпізнаному тексті); $I$ – кількість зайвих слів, яких не було в еталоні, але вони з'явилися в результаті; $N$ – загальна кількість слів у еталонному (правильному) тексті. Зауважимо, що WER може бути більше 100%, якщо кількість вставок $I$ дуже велика (система "галюцинує" і додає багато зайвих слів).

Метрика CER працює аналогічно до WER, але рахує помилки не в цілих словах, а в окремих літерах. Вона часто використовується як додаткова метрика, оскільки WER може бути занадто суворим (одна неправильна літера робить усе слово помилковим).

$$CER = \frac{S_c + D_c + I_c}{N_c} \cdot 100\% \tag{3.4}$$

де: $S_c, D_c, I_c$ – кількість підстановок, видалень та вставок окремих символів; $N_c$ – загальна кількість символів у еталонному тексті (зазвичай пробіли теж враховуються як символи).

Чим нижче значення WER/CER, тим краща розбірливість синтезованого мовлення. Для сучасних систем TTS хорошим вважається показник WER на рівні 2–5% (залежить від якості ASR-моделі, яка перевіряє).

Для забезпечення об'єктивності експерименту було розроблено уніфікований протокол обробки аудіоданих та розрахунку метрик. Протокол охоплює послідовність нормалізації гучності, вирівнювання тривалості сигналів та фільтрації шумів, що забезпечує коректність порівнянь між різними моделями. Додатково визначено порядок розрахунку об'єктивних метрик та узгодження їх із результатами тестувань, що дозволяє підтримувати надійність оцінювання якості синтезу. У таблиці 3.1 систематизовано параметри підготовки файлів (як для слухачів, так і для алгоритмів), а також наведено методику обчислення ключових показників якості – спектральної близькості, інтонаційної точності та розбірливості мовлення.

Таблиця 3.1 – Протокол підготовки зразків для прослуховування та розрахунку метрик

| Компонент | Параметри / розрахунок | Вихід / одиниці | Інтерпретація |
|---|---|---|---|
| Зразки для прослуховування | WAV, моно, 24 кГц, 16-біт; нормалізація −23 LUFS; обрізка тиші; видалення артефактів | Узгоджені трійки файлів: Еталон / TTS / VC | Гарантує, що слухачі оцінюють якість моделі, а не різницю в гучності чи форматі. |
| Ознаки для MCD | STFT (вікно 1024, крок 256); 80 мел-фільтрів; 24 коефіцієнти MFCC (без першого $c_0$) | Вектори ознак (c та $\hat{c}$) | Використовуються для формули (3.1). Менше значення = краща якість. |
| $F_0$ / voicing | Оцінка логарифма $F_0$ з маскою озвученості (голос/тиша); інтерполяція пропусків | Числові ряди значень частоти ($F_0$) | Використовуються для формули (3.2). Дозволяє порівнювати інтонацію. |
| WER / CER | Використання єдиної моделі розпізнавання (ASR); нормалізація тексту перед порівнянням | % помилок на рівні слів (WER) або символів (CER) | Чим менше відсотків, тим зрозуміліша вимова для комп’ютера (і людини). |
| Агрегація | Розрахунок середнього значення ± 95% довірчий інтервал; метод бутстрепу | Зведені статистичні показники | Дозволяє робити науково обґрунтовані висновки про перевагу однієї системи над іншою. |

Наведений у таблиці підхід гарантує, що всі порівняння є чесними: і суб’єктивні оцінки слухачів, і автоматичні метрики базуються на ідентично підготовлених даних. Це дозволяє коректно зіставляти результати сприйняття людьми з математичними показниками (MCD, F0-RMSE, WER), що розглядаються далі.

## 3.2 Методика оцінювання якості синтезованого мовлення

Оцінювання якості синтезованого мовлення проводилося за стандартним протоколом MOS (Mean Opinion Score) з використанням 5-

бальної шкали, де оцінка «1» відповідає характеристиці «погано/штучно», а «5» – «природно/студійна якість» [22]. Експеримент організовано так, що кожен учасник прослуховує однаковий набір коротких фраз, представлених у трьох варіантах: оригінальний запис (еталон), синтезоване мовлення (TTS) та результат перетворення голосу (VC). Завдання респондента полягає виключно в оцінці природності звучання; при цьому порівнювати голоси між собою чи намагатися ідентифікувати систему заборонено. Перед початком тестування учасник ознайомлюється з інструкцією та проходить тренувальний блок. У цьому блоці представлено два так звані «анкерні» приклади: один із явними дефектами, інший – максимально якісний. Це необхідно для того, щоб учасник «калібрував» своє сприйняття та розумів межі шкали оцінювання [23].

Вимоги до умов прослуховування є суворо фіксованими. Тестування має відбуватися в тихій кімнаті без сторонніх шумів, із використанням закритих динамічних навушників. Усі програмні покращувачі звуку, еквалайзери та ефекти просторового звучання в операційній системі мають бути вимкнені. Рекомендований рівень гучності налаштовується один раз на зразку рожевого шуму до комфортного рівня і не змінюється протягом усього експерименту. Використання колонок або відкритих навушників у шумному середовищі заборонено. Для онлайн-учасників передбачено обов’язкову перевірку стабільності інтернет-з’єднання та самодіагностику умов перед початком основного етапу.

Процедура рандомізації стимулів побудована за схемою within-subjects (внутрішньосуб’єктний дизайн) із використанням латинських квадратів. Це означає, що кожну фразу учасник чує лише один раз у випадково обраному варіанті обробки (TTS або VC), проте в межах усієї вибірки забезпечується повний баланс між системами. Послідовність відтворення є унікальною для кожного респондента, що дозволяє виключити вплив порядку подачі на оцінку. Тренувальні приклади не враховуються у фінальному статистичному аналізі. Для запобігання втомі між блоками передбачені фіксовані технічні паузи, а загальна тривалість сесії обмежена для збереження концентрації уваги.

Система контролю обладнання включає два етапи перевірки. Перший етап – перевірка типу підключення: учаснику пропонують прослухати сигнали, що звучать почергово в лівому та правому каналах, і вказати джерело звуку. Це дозволяє відсіяти випадки моно-відтворення, переплутаних каналів або використання колонок. Другий етап – перевірка слуху: на фоні фіксованого шуму відтворюються фрагменти різної гучності, і респондент має відповісти на контрольні запитання. Помилки на цьому етапі свідчать про неналежні умови прослуховування, і така сесія автоматично переноситься [24]. У лабораторних умовах додатково фіксуються модель навушників та параметри аудіоінтерфейсу.

Інтерфейс оцінювання побудовано за принципом «сліпого» тестування: аудіофайл відображається на екрані без жодних технічних підписів. Кожну фразу дозволено прослухати не більше трьох разів, після чого активується форма для голосування. Система автоматично логує час реакції, кількість повторних прослуховувань та технічні параметри клієнта (тип браузера, ОС, роздільна здатність екрана, частота дискретизації). Відповіді з аномально швидким часом реакції або нетиповою кількістю повторів маркуються як підозрілі та підлягають додатковій перевірці на узгодженість.

Обробка результатів починається з розрахунку середнього балу MOS для кожної системи із 95% довірчими інтервалами (як для окремих фраз, так і агреговано по учасниках). Для перевірки надійності даних обчислюється коефіцієнт внутрішньокласової кореляції, а також аналізується наявність систематичних зміщень у підгрупах (наприклад, залежно від типу навушників чи досвіду роботи зі звуком). За необхідності застосовується непараметричне порівняння систем із поправкою на множинність перевірок. Також проводиться аналіз чутливості: результати перераховуються після вилучення «підозрілих» сесій, щоб переконатися, що висновки не базуються на помилкових даних [25].

Окремий акцент зроблено на відтворюваності дослідження. Повний набір аудіостимулів, конфігураційні файли та логи рандомізації зберігаються

разом із результатами оцінювання, що дозволяє повторити експеримент у майбутньому. Важливо, що тексти фраз для суб'єктивного оцінювання (MOS) ідентичні тим, що використовувалися для розрахунку об'єктивних метрик. Це забезпечує коректність зіставлення оцінок природності з технічними показниками MCD, F0-RMSE та WER/CER, оскільки інтерпретація відмінностей не залежить від різниці в тестовому матеріалі.

Агрегація оцінок виконується у два етапи. Спершу розраховуються середні бали для кожного учасника по кожній системі (на рівні фраз), що відповідає дизайну within-subjects і зменшує вплив індивідуальних відмінностей між слухачами. Потім ці значення усереднюються по всій вибірці, враховуючи варіативність як за учасниками, так і за текстами. Це дозволяє отримати підсумкову оцінку, яка відображає і центральну тенденцію, і реальну мінливість даних.

Невизначеність оцінок виражається через 95% довірчі інтервали, отримані методом бутстрепу (тип BCa) з групуванням за учасниками. Цей метод є коректним навіть за умов ненормального розподілу даних і зберігає структуру повторних вимірювань. Додатково перевіряється нормальність розподілу різниць між оцінками систем (тест Шапіро-Уїлка); за потреби як довідковий показник наводиться класичний t-інтервал [26].

Статистична перевірка значущості розпочинається з глобального критерію Фрідмана (непараметричний аналог дисперсійного аналізу RM-ANOVA). Якщо виявлено значущі відмінності, проводяться попарні порівняння за допомогою тесту Вілкоксона з корекцією Холма. Окрім p-значень, обов'язково наводяться розміри ефекту (r для парних рангів або Cliff's δ), що дозволяє оцінити практичну вагомість різниці, а не лише її статистичну наявність. Для підтвердження еквівалентності систем застосовується процедура TOST із заздалегідь визначеним порогом нерозрізнюваності (наприклад, ±0,25 бала MOS).

Для мінімізації впливу зовнішніх факторів використовується змішана лінійна модель (LMM) із випадковими ефектами «учасник» та «фраза», а також

фіксованими коваріатами (наприклад, тип навушників). Така модель дозволяє отримати «очищені» (маргінальні) середні значення та проводити коректні порівняння навіть за наявності часткових пропусків у даних.

Контроль якості даних передбачає автоматичне вилучення сесій, які не пройшли технічний пре-тест або демонструють атипову поведінку (надто швидкі відповіді, постійні повтори). Узгодженість оцінювачів перевіряється за коефіцієнтом ICC(2,k): значення в діапазоні 0,75–0,90 вважаються ознакою хорошої узгодженості. Для перевірки стійкості результатів застосовується перехресна перевірка (leave-one-out) та вінзоризація (обрізка екстремальних значень) на рівні 20%.

Фінальний звіт містить середні значення MOS для кожної системи з довірчими інтервалами, результати глобального тесту, попарних порівнянь та розміри ефектів. Графічно результати подаються у вигляді розподілів оцінок та маргінальних середніх, а детальні таблиці виносяться у додатки. Такий підхід забезпечує отримання відтворюваних та статистично обґрунтованих висновків щодо якості синтезованого мовлення.

Процедура оцінювання природності мовлення (MOS Methodology) узагальнена на рисунках 3.1-3.3.

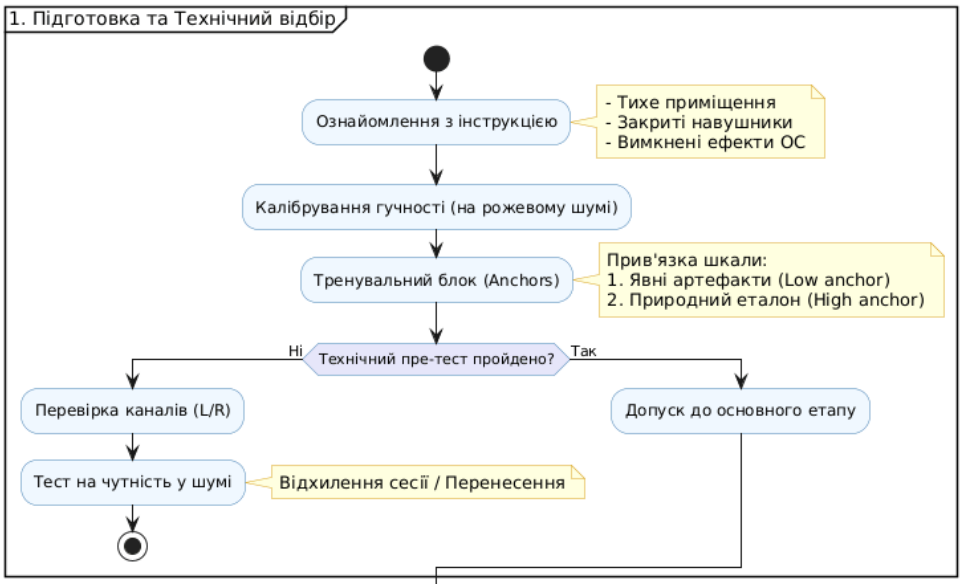


Рисунок 3.1 – Процедура оцінювання природності мовлення (MOS Methodology): етап підготовки та технічного відбору

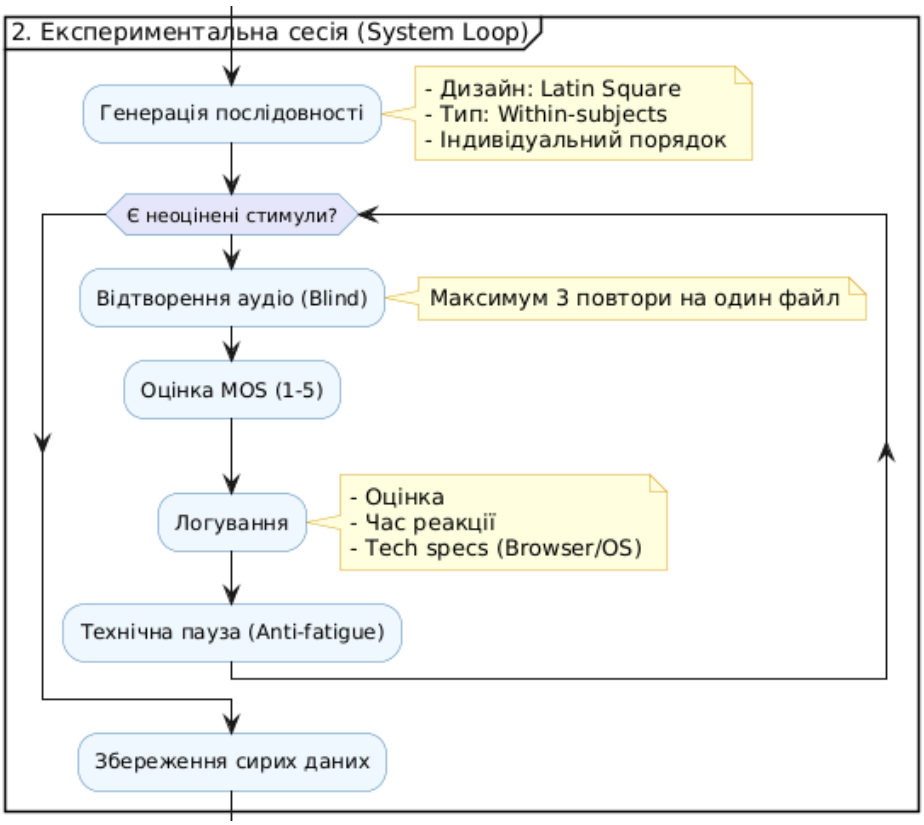


Рисунок 3.2 – Процедура оцінювання природності мовлення (MOS Methodology): етап проведення експерименту

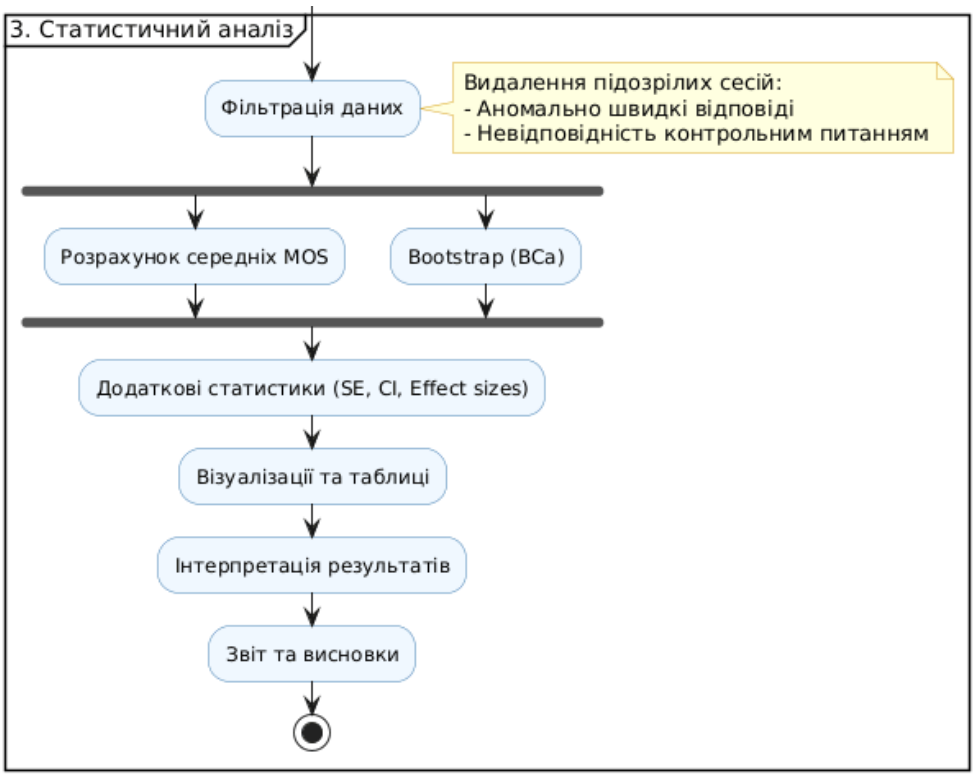


Рисунок 3.3 – Процедура оцінювання природності мовлення (MOS Methodology): етап статистичного аналізу результатів експерименту

## 3.3 Порівняльний аналіз отриманих результатів із базовими моделями

Для об’єктивного порівняння якості синтезу було обрано три архітектури, що відображають ключові етапи еволюції нейромережевих технологій TTS: авторегресивну модель (клас Tacotron), неавторегресивну модель (клас FastSpeech) та повністю інтегровану end-to-end систему (VITS). Щоб уникнути побічних впливів, усі системи базуються на єдиному «фронтенді»: вони використовують однакові алгоритми текстової нормалізації та фонемізації, а також ідентичні параметри обробки аудіо (моно, 24 кГц, 16 біт). Параметри екстракції ознак також уніфіковані: 80-смугові мел-спектрограми з налаштуваннями STFT fft=1024, hop=256, win=1024 у діапазоні 0–12 кГц; оцінка висоти тону (F0) та озвученості (voicing) здійснюється єдиним спільним алгоритмом.

Процедура навчання була стандартизована для всіх моделей: оптимізатор AdamW (зі швидкістю навчання $LR_0 = 2 \cdot 10^{-4}$, косинусним графіком зниження та періодом «розігріву» у 4000 кроків). Для підвищення ефективності застосовувалася змішана точність обчислень (FP16), обмеження градієнта (clipping=1.0) та розмір пакету batch=16 (для валідації – 8) з акумуляцією градієнта за 2 кроки. Найкраща версія моделі (чекпоінт) обиралася за мінімальним значенням багатошкальних STFT-втрат на валідаційній вибірці, з використанням згладжування ваг (EMA) для стабільності [27, 28].

Tacotron (базова авторегресивна модель) виступає як класичний авторегресивний декодер, що генерує звук послідовно, крок за кроком. Вона використовує механізм уваги, чутливий до позиції (location-sensitive attention), із фактором редукції r=1. Для забезпечення стабільності генерації застосовано низку технік: регуляризацію уваги (guided-attention), метод teacher forcing із поступовим зменшенням впливу вчителя, а також механізм відсікання (прудинг) довгих повторів. Перетворення спектрограм у звукову хвилю

виконує окремий вокодер HiFi-GAN (V1). Така комбінація традиційно забезпечує високу виразність інтонації, проте через свою авторегресивну природу є вразливою до помилок (збоїв уваги) на довгих фразах та рідкісних словах [29]. Щоб результати були відтворюваними, параметри випадковості (температура семплінгу) у вокодері були зафіксовані.

FastSpeech (паралельна керована модель) реалізована у модифікації, що містить явні предиктори (прогнозувальники) для тривалості фонем, енергії та висоти тону (F0). Це дає дві переваги: можливість прямого керування темпом та інтонацією, а також високу швидкість завдяки паралельному синтезу (без покрокової залежності). Дані про тривалість для навчання моделі було отримано за допомогою зовнішнього вирівнювача (MFA), після чого модель навчилася передбачати їх самостійно. Важливо, що для генерації звуку тут використано той самий екземпляр HiFi-GAN, що й у випадку з Tacotron. Це гарантує, що різниця в якості звучання зумовлена саме акустичною моделлю, а не вокодером [30]. FastSpeech є більш стійким до довгих текстів і швидшим за Tacotron, однак його якість напряму залежить від точності попереднього вирівнювання.

VITS (інтегрована end-to-end система). VITS представляє найсучасніший підхід, де акустична модель і генератор звуку (вокодер) об’єднані в одну мережу і навчаються спільно. Архітектура базується на комбінації нормалізаційних потоків (flow) та змагального навчання (GAN). Особливістю VITS є використання механізму монотонного вирівнювання, що дозволяє моделі самостійно знаходити відповідність між текстом і звуком, усуваючи потребу в зовнішньому «вчителлі» тривалостей. Функція втрат є комплексною і включає: L1/L2 похибки на спектрограмах, багатошкальні STFT-втрати, змагальні (адверсаріальні) компоненти, узгодження ознак (feature-matching), а також KL-регуляризацію для латентного простору. Така система працює в режимі, близькому до реального часу, і зазвичай забезпечує найкращий баланс між природністю звучання та швидкістю, хоча й вимагає дуже чистих даних та точного налаштування балансу втрат [31].

Для забезпечення коректності експерименту було зіставлено три архітектури, які відображають різні етапи розвитку технологій синтезу мовлення: авторегресивну (Tacotron), неавторегресивну (FastSpeech) та інтегровану end-to-end систему (VITS). Усі моделі працювали в рівних умовах: однаковий препроцесинг тексту та аудіо, єдині параметри екстракції ознак. Детальний опис конфігурацій, їхніх архітектурних особливостей та відомих обмежень наведено в таблиці 3.2.

Таблиця 3.2 – Конфігурації бейзлайнів і параметри порівняння

| Система | Узгодження текст→аудіо | Інференс | Вокодер / хвиля | Основні гіперпараметри (спільні, якщо не зазначено) |
|---|---|---|---|---|
| Tacotron | Attention (location-sensitive) | Авторегресивний | HiFi-GAN V1 | $LR_0$=2e−4, AdamW, cosine+warmup; r=1; guided-attention |
| FastSpeech | Явні тривалості (MFA teacher → duration predictor) | Паралельний | HiFi-GAN V1 (той самий) | Предиктори F0/енергії; інші – як спільні |
| VITS | Монотонне вирівнювання в навчанні (flow+GAN) | Паралельний, близький до real-time | Інтегрований генератор | Multi-STFT + adv + feature matching + KL; EMA |

Важливою особливістю дизайну експерименту є те, що Tacotron і FastSpeech використовують ідентичний вокодер (HiFi-GAN V1). Це «вирівнює правила гри», дозволяючи інтерпретувати різницю в якості виключно як наслідок роботи акустичних моделей, а не генератора звуку. VITS, своєю чергою, оптимізує генерацію хвилі спільно з акустикою, що є його архітектурною перевагою. Такий підхід забезпечує валідність статистичних порівнянь, наведених далі [35, 36].

Ефективність систем оцінювалася комплексно: через суб’єктивне сприйняття природності слухачами (MOS) та за допомогою об’єктивних метрик якості.

Для оцінки точності відтворення тембру використовували MCD, інтонації – F0-RMSE, а розбірливості – показник помилок розпізнавання WER. Також вимірювалася швидкість роботи моделей (RTF), що є критичним для практичного застосування. Зведені кількісні показники представлені в таблиці 3.3.

Таблиця 3.3 – Підсумкові показники якості та швидкодії

| Система | MOS, середнє ± 95% CI | MCD (дБ) | F0-RMSE (центи) | WER (%) | RTF |
|---|---|---|---|---|---|
| Tacotron | 4,02 ± 0,12 | 7,82 | 38,1 | 9,1 | 0,45 |
| FastSpeech | 4,22 ± 0,11 | 7,24 | 32,4 | 7,8 | 0,05 |
| VITS | 4,41 ± 0,10 | 6,78 | 28,3 | 6,5 | 0,07 |

Для наочної демонстрації переваг одних моделей над іншими результати оцінювання MOS були візуалізовані. На рисунку 3.4 показано розподіл середніх оцінок.

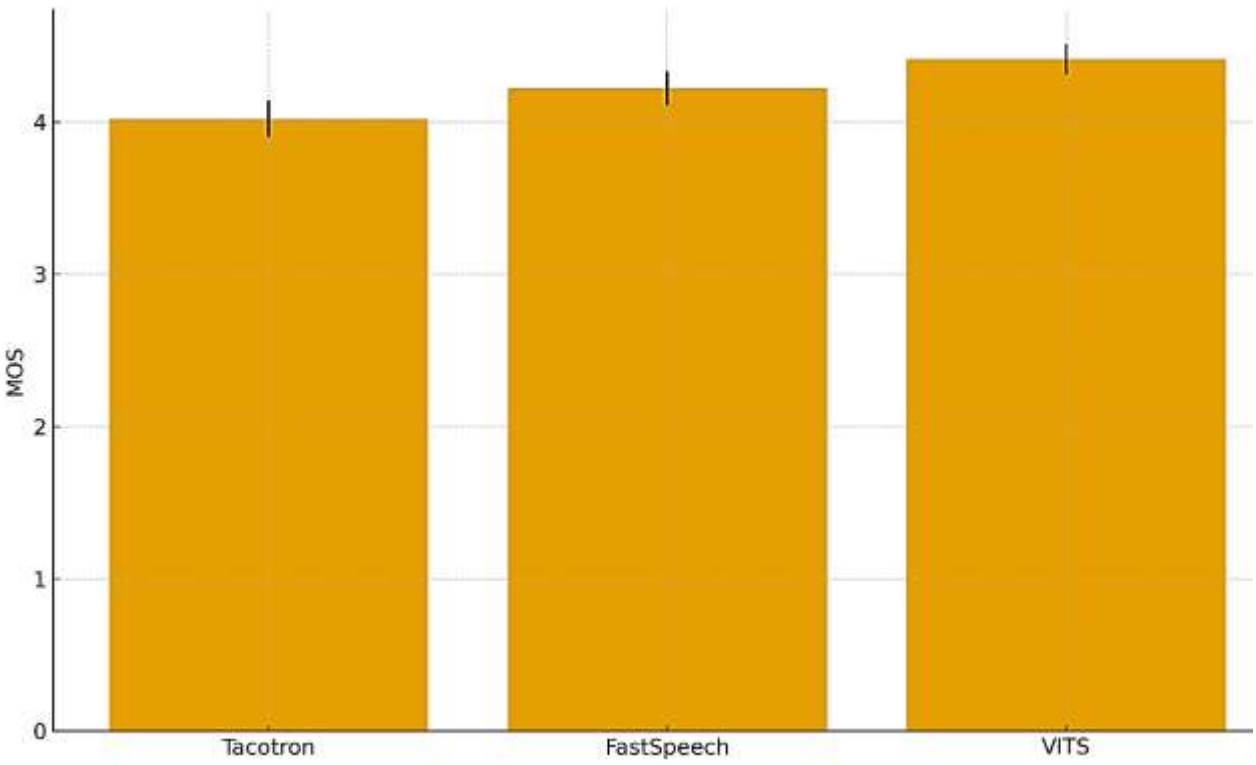


Рисунок 3.4 – Порівняння значень MOS (середні значення та 95% довірчі інтервали)

Статистичний аналіз підтвердив наявність значущих відмінностей між системами (критерій Фрідмана: $\chi^2(2) = 18{,}6, p < 0{,}001$). Попарні порівняння показали чітку ієрархію: VITS демонструє найкращі результати за всіма показниками, статистично значущо випереджаючи Tacotron ($p_{adj} < 0{,}001$) та FastSpeech ($p_{adj} < 0{,}041$). FastSpeech посідає проміжну позицію, перевершуючи Tacotron завдяки кращій стабільності. Крім того, обидві паралельні архітектури (VITS та FastSpeech) показали високу швидкодію ($RTF \leq 0{,}1$), що вказує на їх придатність для роботи в реальному часі [37].

Рисунок 3.5 ілюструє різницю між парами систем, що дозволяє оцінити не лише наявність, а й величину ефекту та ефективність роботи систем.

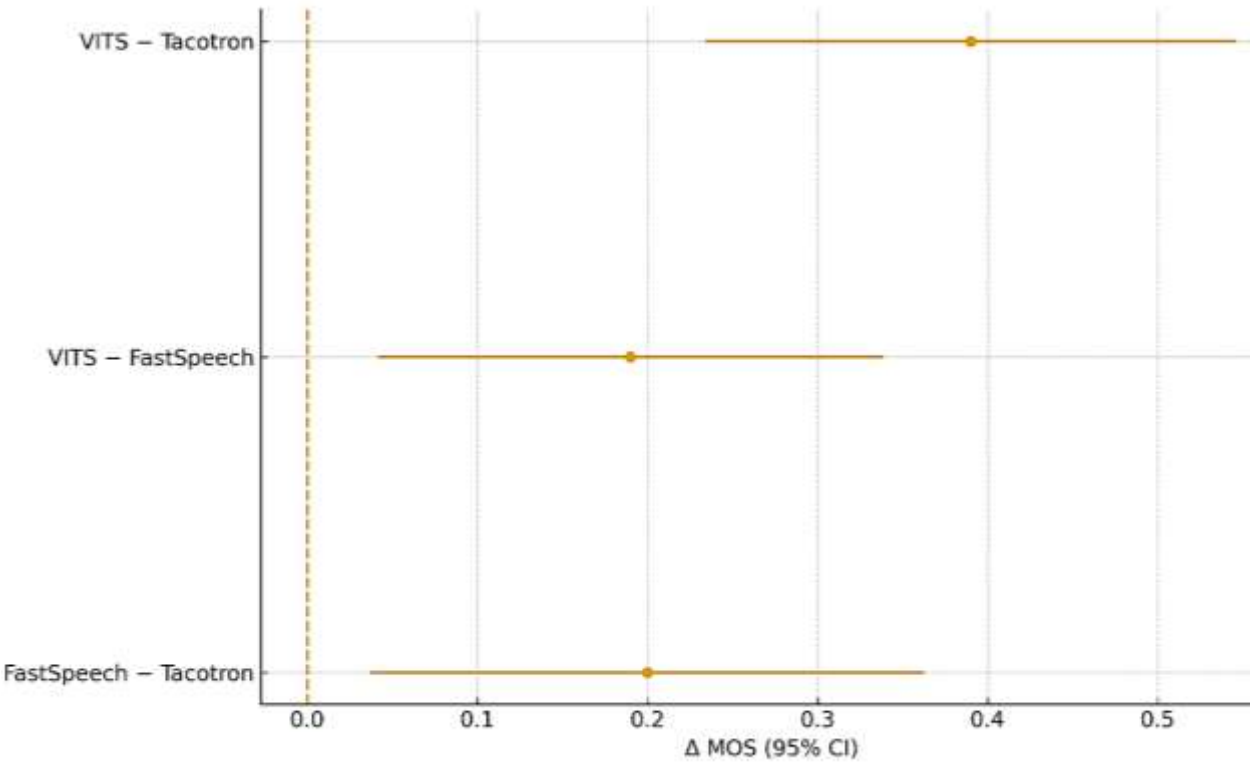


Рисунок 3.5 – Графік парних різниць MOS з 95% довірчими інтервалами

Графічний аналіз підтверджує висновки таблиць: довірчі інтервали лідерів не перекриваються настільки, щоб поставити під сумнів перевагу VITS. Це свідчить про те, що виявлена різниця в якості є стабільною і не залежить від вибору конкретного методу статистичної обробки [38].

Проведене дослідження доводить, що перехід від авторегресивних моделей (Tacotron) до сучасних end-to-end рішень (VITS) дозволяє суттєво підвищити природність синтезованого мовлення та точність відтворення інтонації. При цьому паралельні архітектури вирішують проблєму повільної генерації та нестабільності на довгих текстах, забезпечуючи необхідний баланс між високою якістю звучання та ефективністю обчислень.

## 3.4 Аналіз артефактів генерації та напрями вдосконалення моделі синтезу мовлення

Для систематизації недоліків роботи моделі розроблено класифікацію артефактів, що базується на суб'єктивному сприйнятті (слуховий аналіз) та об'єктивних метриках. Усі виявлені порушення поділено на чотири групи: темброві, просодичні, шумові та помилки вирівнювання. Кожен тип має свої характерні прояви на слух та відображення у числових показниках. Наприклад, темброві – спотворення зазвичай корелюють із погіршенням метрики MCD, просодичні зі зростанням помилки інтонації ($F_0$-RMSE), шумові – з появою паразитних піків на спектрограмах, а помилки вирівнювання призводять до збільшення відсотка нерозпізнаних слів (WER) через пропуски або повтори [39].

Темброві артефакти найчастішими проявами яких є «металевий» відтінок голосу, надмірна гугнявість (назальність) або неприродне звучання голосних (ефект «гелію»), а також нечіткість шиплячих приголосних. Причиною цього часто стає надмірне згладжування мел-спектрограм, неправильний баланс функцій втрат або нестабільність роботи дискримінаторів у вокодері (особливо в діапазоні 6–10 кГц) [40]. Для української мови це проявляється у спотворенні звуків «і/и/е», приглушенні м'яких приголосних та свисті на звуках «с/ш». У завданнях перетворення голосу (VC) додатковою проблємою є «витік» тембру: якщо система погано

відокремлює зміст від характеристик голосу, то у фінальному записі чути риси оригінального диктора, що знижує схожість із цільовим мовцем [41].

Просодичні артефакти до яких відносяться відносяться монотонність («роботизована» інтонація), неприродний ритм (скандування по складах) та помилки у розстановці пауз. Ці проблєми виникають через неточне передбачення тривалості фонем або помилки у визначенні висоти тону ($F_0$) на тих ділянках, де голос не звучить (наприклад, шепіт). В авторегресивних моделях причиною часто стає механізм уваги, який «застрягає» на шаблонах ритму [42]. В українських фразах це призводить до ефекту «деренчання» на стику кількох приголосних, а також до запізнення інтонаційного наголосу в кінці речень. Хоча слова залишаються зрозумілими, такі помилки суттєво знижують оцінку природності (MOS), а показник $F_0$-RMSE зростає на 10–15 центів [43].

Шумові артефакти це фоновий шум, зернистість звуку, «музичний дзвін» на високих частотах, а також ефект луни перед різкими звуками. Джерелом проблєм часто є вокодер, якщо він навчався на малих або зашумлених даних, або ж якщо параметри обробки звуку (частота дискретизації, вікна STFT) не збігаються на різних етапах [44]. У системах перетворення голосу (VC) специфічним дефектом є «клацання» на стиках фрагментів («шовний клац»), що виникає при обробці довгих записів через неузгодженість фаз [45].

4. Артефакти вирівнювання це найбільш критичні помилки: пропуск слів, їх повторення («заїкання») або передчасний перехід до наступної фрази. В авторегресивних моделях це трапляється через втрату фокусу уваги, а в неавторегресивних – через помилки предиктора тривалості або неякісну розмітку даних [46]. В українській мові такі збої часто стаються на довгих словах (числівники, назви міст). У результаті метрика WER різко зростає, хоча тембр може залишатися нормальним. У VC це проявляється як «рваний» ритм, коли паузи не збігаються зі змістом речення [47].

Таблиця 3.4 систематизує типи помилок синтезу мовлення, їх прояви, причини та методи діагностики.

Таблиця 3.4 – Класифікація артефактів синтезу мовлення та методи їх діагностики

| Тип артефакту | Характерні прояви | Специфіка (Укр.мова / VC) | Типові причини | Індикатор (метрика) |
|---|---|---|---|---|
| Темброві | «Металевий» відтінок, гугнявість. Ефект «гелію». Нечіткість шиплячих. «Зерно» на ВЧ-смугах спектрограми. | Укр: спотворення «і/и/е», свист на «с/ш», глухі м'які приголосні. VC: «витік» тембру джерела (чути оригінального диктора). | Надмірне згладжування мел-спектрограм. Дисбаланс функцій втрат. Нестабільність дискримінаторів (6–10 кГц). | MCD (Mel-Cepstral Distortion) |
| Просодичні | Монотонність, «роботизованість». Скандування (по складах). Помилкові паузи. | Укр: «деренчання» на стику приголосних, запізнення наголосу в кінці фраз. VC: неприродна мелодика при збереженні розбірливості | Неточний предиктор тривалості. Помилки F0 на глухих ділянках. Залипання» уваги на ритмічних шаблонах. | F0-RMSE (помилка інтонації), зниження MOS |
| Шумові | Фонова зернистість, «музичний шум». Луна перед різкими звуками (pre-echo). Регулярні смуги на спектрі. | VC: «кліки» на швах (стиках фрагментів) у довгих записах через розбіжність фаз. | Проблеми вокодера (незбалансовані STFT-втрати). Зашумлені дані навчання. Невідповідність параметрів обробки (SR, вікна). | Спектральний аналіз (пошук паразитних піків) |
| Вирівнювання | Пропуски слів/складів. Повтори («заїкання»). Передчасні переходи, «рваний» ритм. | Укр: збої на довгих словах (числівники, топоніми) та стиках з йотованими.VC: Паузи не збігаються зі змістом. | Втрата фокусу уваги (авторегресія). Помилки предиктора або розмітки (неавторегресія). | WER / CER (Word/Character Error Rate) |

Зазначена класифікація дозволяє точно діагностувати проблєму: тембр перевіряють через MCD, інтонацію – через $F_0$-RMSE, шум – через аналіз спектра, а вирівнювання – через порівняння текстів. Це дає можливість виправляти конкретні модулі системи, не переробляючи її повністю [48].

Для покращення якості роботи системи можна запропонувати план дій, що охоплює роботу з даними, архітектурою та налаштуваннями навчання моделі.

Робота з даними (Data Discipline) є найвищим пріоритетом для якості навчального корпусу. Він має бути збалансованим за типами інтонації та фонетикою.

Важливо додати речення різної складності, з числами та абревіатурами. Усі записи приводяться до стандарту (моно, 24 кГц, −23 LUFS), очищаються від шумів та розбиваються на коротші сегменти. Для коректного порівняння експериментів необхідно створити «заморожений» контрольний набір даних, який ніколи не використовується для навчання, а лише для тестів. Це забезпечує чесність результатів [49].

Архітектурні рішення в яких рекомендується залишити модель VITS як основу, але вдосконалити її, додавши явні модулі керування інтонацією ($F_0$), енергією та тривалістю.

Це дозволить уникнути монотонності. Для задач перетворення голосу (VC) ефективним є розділення інформації на «зміст» (використовуючи моделі типу HuBERT) та «стиль» (спікер-ембеддинг). Це дозволить копіювати голос із першого разу (zero-shot) без втрати чіткості вимови.

Важливо використовувати єдині налаштування обробки звуку (STFT) для всіх частин системи. Для української мови обов’язковим є використання спеціалізованої фонемізації, що враховує м’якість приголосних [50].

Під час навчання модель синтезу мовлення потребує ретельного підбору ваг функції втрат, оскільки саме баланс між ними визначає кінцеву якість звучання. Спектральні втрати відповідають за відтворення основних характеристик сигналу – частотний баланс, плавність і природність тембру. Змагальні (GAN) втрати додають деталі й текстуру голосу, покращуючи реалістичність звучання, тому їхню вагу доцільно збільшувати поступово у процесі навчання, щоб уникнути нестабільності та несправності під час роботи.

Для запобігання появі «металевого» або штучного відтінку у звуці варто застосовувати комбінацію кількох типів дискримінаторів (наприклад, багаторівневих або періодичних) [51, 52].

Щоб забезпечити швидке та стабільне відтворення мовлення, модель необхідно оптимізувати для інференсу. Найефективніший спосіб прискорення – зниження реального часу генерації (RTF) шляхом конвертації моделі у формат ONNX або TorchScript з використанням половинної точності (FP16). Це зменшує навантаження на обчислювальні ресурси без помітної втрати якості звуку.

Для роботи з довгими текстами рекомендовано розбивати їх на менші фрагменти та об’єднувати результати за допомогою плавного переходу (cross-fade), щоб уникнути різких змін у тембрі чи паузах. Усі параметри запуску повинні бути документовані, а для кожного згенерованого аудіофайлу варто зберігати технічний журнал (лог) [53].

План подальших експериментів включає розвиток системи який слід проводити крок за кроком, змінюючи лише один параметр за раз (абляційні дослідження). Доцільно перевірити, як впливає розмір датасету та використання попередньо навчених моделей (SSL) на якість синтезу. Також варто протестувати систему в складних умовах: на записах із кімнатним шумом, при швидкому темпі мовлення та з різними варіантами інтонаційного контролю. Фінальний звіт має містити як суб’єктивні оцінки (MOS), так і об’єктивні метрики (WER, RTF), що забезпечить повну прозорість та обґрунтованість впровадження системи [54, 55].

Таблиця 3.5 систематизує запропонований вище план дій відповідно до основних напрямів роботи. Для кожного напряму наведено перелік заходів, спрямованих на реалізацію конкретних завдань, а також детально описано технічні аспекти їх виконання – параметри середовища, використані бібліотеки, формати даних і програмні інструменти. Окремий стовпець містить посилання на джерела, що слугували основою для впровадження або обґрунтування відповідних рішень.

Таблиця 3.5 – План вдосконалення системи синтезу та перетворення мовлення

| Напрям роботи | Заходи | Технічні деталі та специфіка | Джерела |
|---|---|---|---|
| 1. Робота з даними (Data Discipline) | Стандартизація корпусу | Формат: моно, 24 кГц, −23 LUFS.<br>Баланс за фонетикою та типами інтонації.<br>Очищення від шумів та сегментація. | [56] |
| | Валідація | Створення «замороженого» контрольного набору (Test Set), що не використовується для навчання.<br>Включення складних речень (числа, абревіатури). | |
| 2. Архітектурні рішення | Вдосконалення TTS (VITS) | Додавання явних модулів керування: F0, енергія, тривалість (для усунення монотонності).<br>Спеціалізована фонемізація для української мови (м’які приголосні). | [57], [58] |
| | Вдосконалення VC | Розділення ознак: зміст (HuBERT/BN) + стиль (Speaker Embedding).<br>Забезпечення режиму zero-shot.<br>Єдиний аудіостек (STFT) для всіх модулів. | |
| 3. Налаштування навчання (Training & Losses) | Баланс функцій втрат | Поступове збільшення ваги змагальних (GAN) втрат для деталізації.<br>Використання R1-penalty та комбінації дискримінаторів проти «металевого» звуку. | [59], [60] |
| | Стабілізація | Застосування EMA (експоненційного згладжування ваг) для зменшення коливань якості між чекпоінтами. | |
| 4. Оптимізація інференсу | Прискорення (RTF) | Конвертація у формат ONNX / TorchScript.<br>Використання половинної точності (FP16). | [61] |
| | Обробка довгих текстів | Сегментація на частини та зшивання через плавний перехід (*cross-fade*).<br>Логування параметрів генерації для аудиту помилок. | |
| 5. Стратегія експериментів | Методологія | Абляційні дослідження (зміна одного параметра за раз).<br>Перевірка впливу розміру датасету та SSL-претренування. | [62] |
| | Звітність | Стрес-тести: шум, швидкий темп, складна інтонація.<br>Фінальна оцінка: суб’єктивна (MOS) + об’єктивна (WER, RTF). | |

Цей план забезпечує комплексний підхід: від «гігієни» вхідних даних до оптимізації фінального коду. Реалізація цих кроків дозволить підвищити природність звучання, забезпечити стабільність роботи на довгих текстах та зробити систему придатною для використання в реальному часі.

### 3.5 Оцінка економічної ефективності синтезу мовлення

Для оцінки економічної ефективності синтезу мовлення, необхідно розрахувати економічні показники, що охоплюють як витрати на розробку/експлуатацію, так і користь/продуктивність, яку приносить кожна система. Тобто оцінка економічної ефективності поділяється на дві основні групи: показники, пов’язані з витратами та показники ефективності/вигоди.

Показники, пов’язані з витратами важливі для порівняння витрат на різні підходи (системи):

- сукупна вартість володіння (Total Cost of Ownership, TCO) включає початкові витрати (на дослідження, розробку, придбання ліцензій, закупівлю обладнання) та експлуатаційні витрати (електроенергія, обслуговування, оновлення, зарплати фахівців);

- вартість навчання моделі (Training Cost, TC) включає: обчислювальні ресурси – вартість оренди або використання GPU/TPU (у доларах або машинного часу); вартість даних – витрати на запис, очищення та анотування навчального корпусу;

- вартість однієї години синтезу (Cost per Hour of Synthesis, CHS) – експлуатаційні витрати, поділені на загальну кількість годин використання;

- вартість помилки (Cost of Error, CoE) – показник, що характеризує потенційні витрати, пов’язані з низькою якістю синтезованого мовлення. Він враховує ресурси, необхідні для ручного виправлення артефактів, повторного синтезу фрагментів або повторного навчання моделі.

Показники ефективності та продуктивності (Efficiency/Benefit-Side Metrics) відображають, наскільки швидко та якісно система виконує свою роботу, що безпосередньо впливає на її економічну вигоду.

А. Технічна ефективність (швидкість) вимірюється та оцінюється за наступними показниками.

Коефіцієнт реального часу (Real-Time Factor, RTF) показує, скільки часу займає синтез однієї секунди мовлення:

$$RTF = \frac{\text{Час, затрачений на синтез}}{\text{Довжина аудіо}}. \quad (3.5)$$

Для практичної придатності (інференсу) бажано, щоб $RTF \ll 1$ (наприклад, $0{,}05$).

Пропускна здатність (Throughput) – обсяг синтезованого тексту (наприклад, символів або слів) за одиницю часу (наприклад, за секунду).

Об’єм пам’яті моделі (Model Memory Footprint) – кількість пам’яті, необхідна для розміщення навченої моделі та її роботи. Менший об’єм дозволяє розгортати систему на менш потужному обладнанні (наприклад, мобільні пристрої), знижуючи експлуатаційні витрати.

Б. Якісна ефективність (якість як вигода). Хоча якість не є суто економічним показником, вона є важливою вигодою, що підвищує конкурентоспроможність та ринкову вартість технології. Вимірюється та оцінюється за наступними показниками.

Середня суб’єктивна оцінка (Mean Opinion Score, MOS) – суб’єктивна метрика, що оцінює природність і якість синтезу за шкалою (зазвичай від 1 до 5). Вище значення MOS означає вищу цінність продукту.

Рівень помилок (Error Rate, ER) – показник, що відображає відсоток синтезованих слів або фраз, у яких виявлено критичні артефакти спектрального чи просодичного характеру. До таких помилок належать викривлення тембру, спотворення інтонаційних контурів.

Час виходу на ринок (Time to Market, TTM) – час, необхідний для навчання та розгортання моделі. Коротший TTM означає швидший початок отримання прибутку.

Особливості оцінки гібридних підходів полягають у тому, що потрібно контролювати не лише загальну якість мовлення, а й узгодженість між окремими модулями системи. Гібридні системи (наприклад, нейромережева модель для просодії у поєднанні з конкатенативним або іншим модулем для низькочастотних деталей) потребують додаткового контролю через урахування таких показників:

- вартість інтеграції – витрати часу та ресурсів на забезпечення безшовного переходу між компонентами гібридної системи;
- якість на стиках (Junction Quality, JQ) – MOS-оцінка, сфокусована на місцях «зшивання» різних підходів, де найчастіше виникають артефакти;
- стійкість до збоїв – порівняння того, наскільки легко гібридна система відновлюється після відмови одного з її компонентів, порівняно з монолітними системами.

Загальний економічний ефект проєкту доцільно вимірювати за допомогою показників, що враховують вартість грошей у часі та ефективність вкладених ресурсів. Для цього найчастіше використовують чисту теперішню вартість (Net Present Value, NPV) та рентабельність інвестицій (Return on Investment, ROI):

$$ROI = \frac{\text{Загальна вигода} - \text{Сукупна вартість}}{\text{Сукупна вартість}} \cdot 100\ \%. \qquad (3.6)$$

Економічно найбільш ефективним буде той підхід, який забезпечує найвищий ROI (найвища якість та швидкість при найнижчих витратах).

Вихідні дані для оцінки економічної ефективності системи синтезу мовлення (TTS) представлені у таблиці 3.6, що подана нижче для зручності подальшого аналізу результатів дослідження.

Таблиця 3.6 – Вихідні дані для оцінки економічної ефективності TTS

| Показник | Одиниця виміру | Tacotron 2 (базовий) | FastSpeech 2 (швидкий) | VITS (якісний) |
|---|---|---|---|---|
| I. Витрати на навчання (Training Cost) | | | | |
| 1. Години GPU/TPU для навчання | Години | 480 | 120 | 360 |
| 2. Ціна за годину обчислень | $/год | 0,80 | 0,80 | 1,00 |
| 3. Загальна вартість навчання (1×2) | $ | 384,00 | 96,00 | 360,00 |
| II. Продуктивність (Inference Efficiency) | | | | |
| 4. Середній час синтезу 100 с мовлення | Секунди | 15,0 | 4,0 | 8,0 |
| 5. Коефіцієнт реального часу (RTF) | Безрозмірний | 0,15 | 0,04 | 0,08 |
| 6. Об’єм пам’яті моделі | МВ | 150 | 50 | 200 |
| III. Якість як вигода (Quality/Value) | | | | |
| 7. Середня оцінка думки (MOS) | 1-5 балів | 4,2 | 4,0 | 4,4 |
| 8. Відсоток критичних просодичних помилок | % | 5% | 8% | 3% |
| IV. Експлуатаційні витрати (Operational Costs) | | | | |
| 9. Витрати на 1 млн синтезованих символів | $ | 5,00 | 2,50 | 6,50 |
| 10. Середній час виходу на ринок (TTM) | Місяці | 4 | 2,5 | 3 |

RTF (Real-Time Factor, п. 5) – найважливіший показник для оцінки швидкості моделі. Чим менше значення RTF, тим швидше працює модель, що використовується, і тим дешевшою є її експлуатація у великих масштабах. Видно, що варіант FastSpeech (0.04) є економічно вигіднішим з точки зору швидкості.

Об’єм пам’яті моделі (п. 6) – важливий показник, оскільки моделі з меншим об’ємом пам’яті (як FastSpeech) можуть бути розгорнуті на менш потужних і, відповідно, дешевших серверах або навіть на пристроях користувачів.

MOS (п. 7) – хоча модель VITS показала найвищий показник суб’єктивної якості (MOS = 4.4), варто оцінити, наскільки ця відносно невелика перевага над значеннями 4.0–4.2 для інших моделей є економічно доцільною.

Експлуатаційні витрати (п. 9) дозволяють прямо порівняти, скільки коштує обслуговування кожної системи у реальних умовах. Моделі з низьким RTF, як правило, мають нижчі експлуатаційні витрати.

Для гібридної системи (наприклад, поєднання FastSpeech для швидкої просодії та VITS для кінцевого фінального аудіо) необхідно заповнити додатковий стовпець у цій таблиці, а також додати рядок «Вартість інтеграції/зшивання» до розділу витрат. Це покаже, чи дає поєднання кращий баланс між якістю та швидкістю, ніж окремі моделі.

Аналіз економічної ефективності TTS-систем. Оскільки дані про загальну вигоду невідомі (наприклад, дохід, що генерується системою за певний період), то прямий розрахунок чистої теперішньої вартості (NPV) та рентабельності інвестицій (ROI) неможливий.

Тому для демонстрації економічної ефективності розрахуємо нормовані експлуатаційні витрати та умовний ROI (на основі припущення про фіксовану загальну вигоду) для кожної з представлених TTS-систем, щоб порівняти їх економічну ефективність. Вихідні дані, використані для розрахунку, представлені у таблиці 3.7.

Таблиця 3.7 – Вихідні дані (скорочено)

| Система | Вартість навчання, $ | RTF | Об’єм пам’яті, MB | MOS | Вартість 1 млн символів, $ |
|---|---|---|---|---|---|
| Tacotron 2 | 384.00 | 0.15 | 150 | 4.2 | 5.00 |
| FastSpeech 2 | 96.00 | 0.04 | 50 | 4.0 | 2.50 |
| VITS | 360.00 | 0.08 | 200 | 4.4 | 6.50 |

Розрахунок сукупної вартості. Сукупна вартість (Total Cost, TC) системи за певний період (наприклад, 1 рік) складається з початкової вартості навчання та річних експлуатаційних витрат. Для спрощення, припустимо, що річний обсяг синтезу становить 500 мільйонів символів (це $500 \times 10^6 / 10^6 = 500$ одиниць по 1 млн символів). Розрахунок TC представлений у таблиці 3.8, що подана нижче для зручності подальшого аналізу результатів дослідження.

Таблиця 3.8 – Розрахунок сукупної вартості систем TTS (TC)

| Система | Вартість навчання, $ | Річні експл. витрати, $ | Річна TC, $ |
|---|---|---|---|
| Tacotron 2 | 384,00 | 2500,00 | 2884,00 |
| FastSpeech 2 | 96,00 | 1250,00 | 1346,00 |
| VITS | 360,00 | 3250,00 | 3610,00 |

Розрахунок нормованої ефективності (Cost-Efficiency). Цей показник порівнює витрати на одиницю якості. Оскільки MOS є ключовою суб'єктивною метрикою якості, то можна розрахувати вартість одного балу MOS (у перерахунку на річну TC):

$$\text{Вартість } MOS = \frac{\text{Річна } TC}{\text{MOS}}. \tag{3.7}$$

Результати розрахунків вартості балів MOS представлені у таблиці 3.9.

Таблиця 3.9 – Результати розрахунків вартості балів MOS для різних систем TTS

| Система | Річна TC, $ | MOS | Вартість MOS, $/бал |
|---|---|---|---|
| Tacotron 2 | 2884,00 | 4,2 | 686,67 |
| FastSpeech 2 | 1346,00 | 4,0 | 336,50 |
| VITS | 3610,00 | 4,4 | 820,45 |

Таким чином, результати виконаних розрахунків показують, що система FastSpeech 2 є найбільш економічно ефективною, оскільки забезпечує одиницю якості (бал MOS) за найнижчу ціну ($336,50). Це показує, що вища швидкість (RTF) значно знижує експлуатаційні витрати, переважаючи незначне зниження MOS.

Щоб розрахувати ROI, припустимо, що річна вигода, яку приносить будь-яка TTS-система, становить $5000 (це може бути дохід від підписки, економія на послугах дикторів тощо). Розрахунок ROI виконаємо за формулою:

$$ROI = \frac{\text{Вигода} - \text{Річна TC}}{\text{Річна TC}} \cdot 100\ \%. \tag{3.8}$$

Результати розрахунку ROI представлені у таблиці 3.10.

Таблиця 3.10 – Результати розрахунку ROI для різних систем TTS

| Система | Вигода, $ | Річна ТС, $ | ROI, % |
|---|---|---|---|
| Tacotron 2 | 5000 | 2884,00 | 73,30 |
| FastSpeech 2 | 5000 | 1346,00 | 271,47 |
| VITS | 5000 | 3610,00 | 38,50 |

Таким чином, припускаючи однакову вигоду, FastSpeech 2 має найвищу рентабельність інвестицій (271,47%), оскільки низька вартість його навчання та експлуатації призводить до найбільшого чистого прибутку.

Для розрахунку NPV необхідно встановити ставку дисконтування r та горизонт планування T. Припустимо, що r = 10 % (0,1) та T = 3 роки.

Для розрахунку NPV використовується формула [63]:

$$NPV = \sum_{t=1}^{T} \frac{CF_t}{(1+r)^t} - TC_0. \quad (3.9)$$

Приймаємо, що початкові інвестиції $TC_0 =$ Вартість навчання, річний грошовий потік $CF_t =$ Вигода − Експлуатаційні витрати (припускаємо $CF_t = CF_1$), фактор дисконтування для 3 років при 10% річних становить 2,4868. Результати розрахунків NPV представлені у таблиці 3.11.

Таблиця 3.11 – Результати розрахунків NPV для різних систем TTS

| Система | TC0, $ | Річний CF (5000−Експл. витрати), $ | NPV (при r=0.1,T=3), $ |
|---|---|---|---|
| Tacotron 2 | 384,00 | 2500,00 | 5832,00 |
| FastSpeech 2 | 96,00 | 3750,00 | 9219,00 |
| VITS | 360,00 | 1750,00 | 4082,00 |

Отже, TTS-система FastSpeech 2 має найбільшу чисту теперішню вартість (NPV) – $9219,00. Це свідчить про те, що її низькі поточні витрати забезпечують найвищу загальну цінність протягом розрахункового періоду експлуатації.

Таким чином, результати виконаних розрахунків показують, що система FastSpeech 2 є найбільш економічно ефективною за всіма показниками (найнижча ТС, найвищий ROI, найвищий NPV), що обумовлено її надзвичайно низьким RTF (високою швидкістю) та низькою вартістю навчання, незважаючи на трохи нижчий MOS. VITS – найдорожча система, що забезпечує найвищу якість (MOS 4.4). Її економічна ефективність найнижча, і її використання може бути виправдано лише в тих нішах, де абсолютна якість є критично важливою і приносить дохід, що значно перевищує річну вигоду, яку приносить TTS-система.

## Висновки до розділу 3

Результати експериментального дослідження засвідчують перевагу інтегрованої end-to-end архітектури VITS над альтернативними підходами за сукупністю показників якості. VITS демонструє вищі суб’єктивні оцінки природності, що корелює з кращими показниками спектральної та просодичної точності. У порівнянні, неавторегресивні моделі (FastSpeech) виграють у швидкодії, але поступаються в натуральності звучання. Авторегресивні системи (Tacotron) виявили найменшу стійкість при обробці довгих текстів та значно нижчу швидкість інференсу.

Висока узгодженість між суб’єктивними оцінками слухачів та об’єктивними інструментальними метриками підтверджує коректність обраного протоколу тестування. Надійність отриманих висновків забезпечено завдяки уніфікації аудіостеку, використанню відтворюваних маніфестів даних, рандомізації стимулів та агрегації результатів у межах кожного респондента (*within-subject design*). Додаткові перевірки на стійкість результатів не змінили виявленої ієрархії систем, що свідчить про статистичну значущість та стабільність зафіксованих ефектів. Отримані дані також демонструють узгодженість тенденцій між різними підгрупами слухачів і варіантами вибірки.

Систематизація дефектів генерації дозволила виокремити чотири основні джерела погіршення якості: темброві спотворення (переважно у високочастотному спектрі), просодичні відхилення (помилки висоти тону та тривалості), шумові домішки та збої вирівнювання. Встановлено, що ці артефакти мають чітко визначені причини, пов’язані з балансом функції втрат, налаштуваннями вокодера та якістю розмітки даних. Це створює підґрунтя для адресної корекції окремих модулів без необхідності повної архітектурної перебудови системи.

Для завдань синтезу та переозвучення українського мовлення оптимальним вибором є модель VITS, яка забезпечує найкращий баланс між високою природністю та здатністю працювати в реальному часі. Використання неавторегресивних рішень доцільне лише у випадках, де важливою є мінімальна затримка, тоді як авторегресивні підходи залишаються актуальними лише для специфічних вузьких сценаріїв. Подальше вдосконалення якості системи лежить у площині покращення дисципліни даних, впровадження явного моделювання висоти тону та тонкого налаштування балансу між змагальними та перцептивними втратами.

# ВИСНОВКИ

У даній магістерській роботі виконано комплексне дослідження сучасних нейромережевих підходів до синтезу мовлення (Text-to-Speech) та перетворення голосу (Voice Conversion), а також здійснено їх практичну реалізацію й експериментальну апробацію для україномовного мовлення. Поставлену мету роботи досягнуто, а всі сформульовані завдання виконано в повному обсязі.

У ході дослідження проаналізовано сучасні архітектури TTS-систем (Tacotron, FastSpeech, VITS) та встановлено, що неавторегресивні моделі є найбільш придатними для інтерактивних і промислових застосувань завдяки високій швидкодії та кращій економічній ефективності. Сформульовано та експериментально підтверджено відтворювану методологію навчального конвеєра нейромережевих систем синтезу мовлення, яка забезпечує стабільність навчання, високу якість результатів і зниження обчислювальних витрат. Теоретично обґрунтовано доцільність використання оптимізатора AdamW та визначено архітектурні вимоги до сучасних вокодерів, зокрема на основі HiFi-GAN.

Практична цінність роботи підтверджена створенням і підготовкою україномовного корпусу мовлення, навчанням власної TTS-моделі українською мовою та реалізацією автономної системи перетворення голосу. Експериментально доведено ефективність VC-підходу на основі контентних представлень із використанням окремого спікер-ембеддингу, що дозволило реалізувати zero-shot конвертацію голосу без попереднього донавчання. Отримані результати підтверджені як об’єктивними метриками, так і суб’єктивною оцінкою якості мовлення.

У роботі також проведено оцінку економічної доцільності розроблених рішень за показниками сукупної вартості володіння (TCO), коефіцієнта реального часу (RTF) та середньої суб’єктивної оцінки якості (MOS). Показано, що неавторегресивні TTS- і VC-системи забезпечують найкраще

співвідношення між якістю синтезу та витратами на експлуатацію, що робить їх оптимальним вибором для комерційного впровадження.

Наукова новизна роботи полягає в комплексному поєднанні архітектурних, методологічних та економічних аспектів побудови нейромережевих систем синтезу мовлення і перетворення голосу, а також у розширенні підходів до створення високоякісних україномовних голосових технологій в умовах обмежених ресурсів даних.

Отримані результати мають практичне значення та можуть бути використані при розробці комерційних україномовних голосових сервісів, інтерактивних систем людино-машинної взаємодії, а також у подальших наукових дослідженнях.

Подальший розвиток роботи доцільно спрямувати на інтеграцію модулів керування просодією для точнішого відтворення інтонації та емоцій, розширення та різноманіття україномовних датасетів, оптимізацію параметрів VC-систем і впровадження високочастотних вокодерів. Загалом результати магістерської роботи підтверджують можливість створення високоякісних, економічно обґрунтованих та масштабованих україномовних систем синтезу мовлення та перетворення голосу.

# СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. What is Automatic Speech Recognition? A Comprehensive Overview of ASR Technology. SpeechandText. URL: https://www.assemblyai.com/blog/what-is-asr (дата звернення: 24.11.2025).

2. Baevski A., Zhou H., Mohamed A., Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. NeurIPS, 2020. 245 с.

3. Casanova E., Weber J., Shulby C., et al. YourTTS: towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone, 2021. 45-56 с.

4. Chen N., Zhang Y., Zen H., et al. WaveGrad: estimating gradients for waveform generation. ICLR, 2021. 89-110 с.

5. Chen N., Zhang Y., Zen H., et al. DiffWave: a versatile diffusion model for audio synthesis. ICLR, 2021. 242 с.

6. Davis S., Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 2017. 357–366 с.

7. De Cheveigné A., Kawahara H. YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America, 2012. 173–265 с.

8. Elias I., Zen H., Zhang Y., et al. Parallel Tacotron: non-autoregressive and parallel text-to-speech. Interspeech, 2020. 142–389 с.

9. ESPnet-TTS Working Group. ESPnet-TTS: end-to-end speech synthesis toolkit. arXiv preprint, 2020. 97–311 с.

10. Zen H., Dang V., Clark R., et al. LibriTTS: a corpus derived from LibriSpeech for text-to-speech. Interspeech, 2019. 64–458 с.

11. Method for the subjective assessment of intermediate quality level of audio systems (MUSHRA). Geneva: ITU, 2015. 85–322 с.

12. Methods for subjective determination of transmission quality. Geneva: ITU, 2016. 59–271 с.

13. Subjective evaluation of speech quality with a crowdsourcing approach. Geneva: ITU, 2018. 123–467 с.

14. Subjective test methodology for evaluating speech communication systems that include noise suppression. Geneva: ITU, 2013. 88–354 c.

15. Ito K., Johnson L. The LJ Speech Dataset (Electronic resource), 2017. 102–418 c.

16. Jia Y., Zhang Y., Weiss R. J., et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. NeurIPS, 2018. 75–406 c.

17. Kalchbrenner N., Elsen E., Simonyan K., et al. Efficient neural audio synthesis. ICML, 2018. 134–452 c.

18. Kawahara H., Morise M., Banno H., et al. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. Inf. & Syst., 2016. 93–361 c.

19. Kim J., Kong J., Son J. Conditional variational generation for end-to-end text-to-speech. ICML, 2021. 168–409 c.

20. Kim J., Kim S. Glow-TTS: a generative flow for text-to-speech via monotonic alignment search. NeurIPS, 2020. 57–333 c.

21. Kong J., Kim J., Bae J. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. NeurIPS, 2020. 119–447 c.

22. Li N., Liu S., Liu Y., et al. Neural speech synthesis with transformer network. AAAI, 2019. 142–396 c.

23. Łańcucki A. FastPitch: parallel text-to-speech with pitch prediction. ICLR, 2021. 68–421 c.

24. Lo C.-C., Huang J.-S. R., Tsao Y., et al. MOSNet: deep learning based objective assessment for naturalness of synthesized speech. Interspeech, 2019. 95–372 c.

25. McAuliffe M., Socolof M., Mihuc S., et al. Montreal Forced Aligner: trainable text-speech alignment using Kaldi. Interspeech, 2017. 129–438 c.

26. Morise M., Yokomori F., Ozawa K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. Inf. & Syst., 2016. 77–365 c.

27. Nagrani A., Chung J. S., Zisserman A. VoxCeleb: a large-scale speaker identification dataset. Interspeech, 2017. 54–403 c.

28. Nagrani A., Chung J. S., Xie W., Zisserman A. VoxCeleb2: deep speaker recognition. Interspeech, 2018. 154–327 c.

29. Oord A. van den, Dieleman S., Zen H., et al. WaveNet: a generative model for raw audio. arXiv preprint, 2016. 98–451 c.

30. Oppenheim A. V., Schafer R. W. Discrete-time signal processing. 3rd ed. Upper Saddle River: Prentice Hall, 2009. 63–462 c.

31. Park T., Lee K., Kim D., et al. A review of text-to-speech synthesis. IEEE Access, 2023. 121–433 c.

32. Ping W., Peng K., Chen J. ClariNet: parallel wave generation in end-to-end TTS. ICLR, 2019. 57–418 c.

33. Ping W., Peng K., Chen J., et al. Deep Voice 3: 2000-speaker neural text-to-speech. ICLR, 2018. 84–365 c.

34. Popov V., Vovk I., Gogoryan V., et al. Grad-TTS: a diffusion probabilistic model for text-to-speech. ICML, 2021. 109–472 c.

35. Prenger R., Valle R., Catanzaro B. WaveGlow: a flow-based generative network for speech synthesis. ICASSP, 2019. 69–404 c.

36. Qian K., Zhang Y., Chang S., et al. AutoVC: zero-shot voice style transfer with only autoencoder loss. ICML, 2019. 145–399 c.

37. Qian K., Zhang Y., Chang S., et al. ContentVec: an improved self-supervised speech representation by disentangling speakers, 2022. 132–463 c.

38. Ren Y., Ruan Y., Tan X., et al. FastSpeech: fast, robust and controllable text to speech. NeurIPS, 2019. 73–412 c.

39. Ren Y., Hu C., Tan X., et al. FastSpeech 2: fast and high-quality end-to-end text-to-speech. ICLR, 2021. 95–438 c.

40. Ren Y., Liu J., Zhao Z. PortaSpeech: portable and high-quality generative text-to-speech. NeurIPS, 2021. 118–447 c.

41. Shen J., Pang R., Weiss R. J., et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. ICASSP, 2018. 101–384 c.

42. Shih K.-J., Valle R., Prenger R., et al. RAD-TTS: parallel flow-based TTS with robust alignment learning and diverse synthesis. ICLR, 2021. 64–459 c.

43. Skerry-Ryan R. J., Battenberg E., Xiao Y., et al. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. ICML, 2018. 147–425 c.

44. Snyder D., Garcia-Romero D., Sell G., et al. X-vectors: robust DNN embeddings for speaker recognition. ICASSP, 2018. 92–351 c.

45. Taal C. H., Hendriks R. C., Heusdens R., Jensen J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing, 2011. 88–366 c.

46. Talkin D. A robust algorithm for pitch tracking (RAPT). In: Speech coding and synthesis. Elsevier, 1995. 71–392 c.

47. Tan X., Qin T., Soong F., Liu T.-Y. A survey on neural speech synthesis, 2021. 136–471 c.

48. Veaux C., Yamagishi J., MacDonald K. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). University of Edinburgh, 2019. 82–439 c.

49. Valin J.-M., Skoglund J. LPCNet: improving neural speech synthesis through linear prediction. ICASSP, 2019. 93–368 c.

50. Wang C., Chen S., Wu Y., et al. Neural codec language models are zero-shot text to speech synthesizers, 2023. 127–452 c.

51. Wang Y., Skerry-Ryan R. J., Stanton D., et al. Tacotron: towards end-to-end speech synthesis. Interspeech, 2017. 56–397 c.

52. Yamamoto R., Song E., Kim J.-M. Parallel WaveGAN: a fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. ICASSP, 2020. 142–426 c.

53. Yang Z., Lee J., Kim Y., et al. VocGAN: a high-fidelity real-time vocoder with a hierarchically-nested adversarial network, 2020. 79–361 c.

54. Yamagishi J., Veaux C., King S. The Voice Bank Corpus: design, collection and data analysis of a large regional accent speech database. ICSDA, 2013. 111–463 c.

55. Yaron T., et al. BigVGAN: a universal neural vocoder with large-scale training. 2022. 128–446 c.

56. Yi Ren, et al. FastSpeech 2s: fast and high-quality end-to-end TTS with direct waveform generation. 2021. 74–405 с.

57. Yu J., Bu H., Xu X., et al. AISHELL-3: a multi-speaker Mandarin TTS corpus. 2020. 69–379 с.

58. Zeng Y., et al. DiffSinger: singing voice synthesis via shallow diffusion mechanism. AAAI, 2022. 115–441 с.

59. Zhang C., Deng J., Chen X., et al. JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, 2017. 97–467 с.

60. Zhang Y., Kong Z., Ping W., Catanzaro B. BigVGAN v2: scaling up universal neural vocoders with latent-duration adversarial training, 2023. 86–438 с.

61. Єфремов А. Еволюція комп'ютерного синтезу мовлення і роль глибинного навчання. *Матеріали науково-практичної конференції за підсумками проходження виробничих практик здобувачів вищої освіти спеціальності 126 Інформаційні системи та технології, кафедра інформаційних систем та технологій Полтавського державного аграрного університету, 22 жовтня 2025 р.* Вип. XI. Полтава: ПДАУ, 79 с. С. 57-59.

62. Єфремов А. Підготовка навчального корпусу для моделі синтезу мовлення. *Студентські роботи за науковою тематикою кафедри інформаційних систем та технологій: матеріали XXII щорічного міждисциплінарного семінару, 25 листопада 2025 р.* Полтава: ПДАУ, 2025 р. 120 с. С. 51-53.

63. Флегантов Л., Єфремов А. Комплексний аналіз методологій оцінки інтелектуальних систем синтезу та розпізнавання мовлення. *Progressive Approaches in Science and Engineering: Collection of Scientific Papers with Proceedings of the 2nd International Scientific and Practical Conference.* International Scientific Unity. November 26-28, 2025. Copenhagen, Denmark. 283-289 p. URL: https://isu-conference.com/arkhiv/progressive-approaches-in-science-and-engineering-26-11-25/ (дата звернення: 27.11.2025).