Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)
Кафедра систем штучного інтелекту та аналізу даних
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістр
(назва освітнього ступеня)

на тему: Метод і засіб узгодження потоку латентного простору DDSP-WORLD вокодера

Виконав: студент 6 курсу, групи САм-61
спеціальності 124 «Системний аналіз»
(шифр і назва спеціальності)

(підпис) Кліщ М.В. (прізвище та ініціали)

Керівник (підпис) Яцишин В.В. (прізвище та ініціали)

Нормоконтроль (підпис) Гладьо С.В. (прізвище та ініціали)

Завідувач кафедри (підпис) Яцишин В.В. (прізвище та ініціали)

Рецензент (підпис) (прізвище та ініціали)

Тернопіль
2025

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра Кафедра систем штучного інтелекту та аналізу даних
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри

_____ _____
(підпис) (прізвище та ініціали)

« » 20___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня магістр
(назва освітнього ступеня)

за спеціальністю 124 «Системний аналіз»
(шифр і назва спеціальності)

студенту Кліщу Максиму Володимировичу
(прізвище, ім'я, по батькові)

1. Тема роботи Метод і засіб узгодження потоку латентного простору DDSP-WORLD вокодера

Керівник роботи Яцишин Василь Володимирович, к.т.н., зав. каф. СА
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від «*07*» *11* 2025 року № *4/7-951*

2. Термін подання студентом завершеної роботи

3. Вихідні дані до роботи Літературні джерела, датасет «Tohoku Kiritan singing database»

4. Зміст роботи (перелік питань, які потрібно розробити)
1. Огляд сучасних методів синтезу співочого мовлення
2. Методи та архітектура моделі синтезу співочого голосу на основі узгодження потоку.
3. Експериментальне дослідження та аналіз результатів

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
1. Титульний слайд. 2. Завдання роботи. 3. Мета роботи. 4. Задача синтезу співочого мовлення. 5. Пропонована архітектура моделі. 6. Диференційована цифрова обробка Сигналів. 7. Залишкове векторне квантування. 8. GAN. 9. Узгодження потоку. 10. Тренувальний критерій. 11. Умови експерименту. 12. Якісна оцінка. 13. Оцінка ефективності. 14. Абляційне дослідження. 15. Візуалізація згенерованих результатів моделі. 16. Висновки.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	Окіпний І.Б., зав. кафедри МТ		
Безпека в надзвичайних	Стручок В.С., ст. вик. каф. ОХ		
ситуаціях			

7. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення з завданням		
2.	Підбір джерел про синтез співочого мовлення		
3.	Опрацювання джерел		
4.	Формування методи оцінювання результату синтезу		
	співочого мовлення		
5.	Розробка архітектури моделі синтезу співочого		
	мовлення		
6.	Оформлення розділу «Огляд сучасних методів		
	синтезу співочого мовлення»		
7.	Оформлення розділу «Методи та архітектура моделі		
	синтезу співочого голосу на основі узгодження		
	потоку»		
8.	Оформлення розділу «Експериментальне		
	дослідження та аналіз результатів»		
9.	Виконання завдання до підрозділу «Безпека		
	життєдіяльності»		
10.	Виконання завдання до підрозділу «Основи охорони		
	праці»		
11.	Оформлення кваліфікаційної роботи		
12.	Нормоконтроль		
13.	Перевірка на плагіат		
14.	Попередній захист кваліфікаційної роботи		
15.	Захист кваліфікаційної роботи		

Студент _____ Кліщ М.В.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Яцишин В.В.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Метод і засіб узгодження потоку латентного простору DDSP-WORLD вокодера // Кваліфікаційна робота освітнього рівня «Магістр» // Кліщ Максим Володимирович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра систем штучного інтелекту та аналізу даних, група САм-61 // Тернопіль, 2025 // С. – 73, рис. – 13, табл. – 4, слайдів – 17, додат. – 1, бібліогр. – 64.

Ключові слова: машинне навчання, глибинне навчання, синтез співочого мовлення, узгодження потоку, диференційовна цифрова обробка сигналів, залишкове векторне квантування.

Кваліфікаційна робота присвячена розробці методу синтезу співочого голосу на основі глибинного навчання. Основну увагу зосереджено на поєднанні генеративного моделювання в латентному просторі з диференційовною цифровою обробкою аудіосигналів з метою досягнення високої якості синтезу за обмежених обчислювальних ресурсів.

У першому розділі кваліфікаційної роботи виконано аналітичний огляд задач синтезу мовлення та синтезу співочого мовлення. Розглянуто сучасні моделі синтезу співочого голосу, зокрема підходи на основі варіаційних, дифузійних та потокових моделей. Проаналізовано сучасні TTS-моделі, що використовують узгодження потоку, а також існуючі вокодери. Окрему увагу приділено методам теселяції латентного простору та залишковому векторному квантуванню. Обгрунтовано актуальність досліджуваної задачі.

У другому розділі кваліфікаційної роботи викладено теоретичні основи запропонованого підходу. Описано метод узгодження потоку векторного поля, його модифіковані функції втрат та умовні варіанти, а також механізми узгодження умовних ознак. Наведено архітектуру генеративної моделі та аудіо-автокодера, що включає енкодер, залишкове векторне квантування

та декодер із DDSP-WORLD вокодером. Також розглянуто архітектуру змагально-генеративної моделі та модель ознак музичної партитури, зокрема енкодери фонемних і нотних ознак, модель часового зсуву та модель тривалості.

У третьому розділі кваліфікаційної роботи описано практичну реалізацію запропонованого методу. Наведено характеристики використаного набору даних та процес оптимізації параметрів моделі. Проведено аналіз експериментальних результатів, абляційне оцінювання окремих компонентів архітектури та узагальнений аналіз отриманих результатів. Окремий підрозділ присвячено візуалізації роботи моделі та аналізу роботи енкодера частоти основного тону.

У четвертому розділі кваліфікаційної роботи розглянуто питання охорони праці та безпеки в надзвичайних ситуаціях. Проаналізовано вплив шуму, ультразвуку та інфразвуку на організм людини та наведено засоби захисту від шкідливої дії акустичних факторів. Також досліджено вплив електромагнітного імпульсу на елементи DDSP-WORLD вокодера, розглянуто нормативно-правове забезпечення та методи підвищення стійкості системи до імпульсних збурень.

ANNOTATION

Method and Tool for Latent Space Flow Matching of a DDSP-WORLD Vocoder // Master's Qualification Thesis // Klishch Maksym Volodymyrovych // Ternopil Ivan Puluj National Technical University, Faculty of Computer Information Systems and Software Engineering, Department of Artificial Intelligence Systems and Data Analysis, Group SAm-61 // Ternopil, 2025 // pp. – 73, fig. – 13, tabl. – 4, chair. – 17, annexes. – 1, references – 64.

Keywords: machine learning, deep learning, singing voice synthesis, flow matching, differentiable digital signal processing, residual vector quantization.

The master's qualification thesis is devoted to the development of a singing voice synthesis method based on deep learning. The main focus is placed on combining latent-space generative modeling with differentiable digital signal processing of audio signals in order to achieve high synthesis quality under limited computational resources.

The first chapter of the thesis provides an analytical review of speech synthesis and singing voice synthesis tasks. Modern singing voice synthesis models are examined, including approaches based on variational, diffusion, and flow-based models. Contemporary TTS models employing flow matching, as well as existing vocoders, are analyzed. Special attention is given to methods of latent space tessellation and residual vector quantization. The relevance of the research problem is substantiated.

The second chapter presents the theoretical foundations of the proposed approach. The method of vector field flow matching, its modified loss functions, and conditional variants are described, along with mechanisms for conditioning feature alignment. The architecture of the generative model and the audio autoencoder is presented, including the encoder, residual vector quantization, and a decoder with a DDSP-WORLD vocoder. In addition, the architecture of the adversarial generative model and the musical score feature model are discussed, including phoneme and

note encoders, the time-lag model, and the duration model.

The third chapter describes the practical implementation of the proposed method. The characteristics of the dataset used and the model parameter optimization process are presented. An analysis of experimental results is conducted, including ablation studies of individual architectural components and a comprehensive evaluation of the obtained results. A separate subsection is devoted to the visualization of the model's operation and the analysis of the fundamental frequency encoder.

The fourth chapter addresses occupational safety and emergency safety issues. The effects of noise, ultrasound, and infrasound on the human body are analyzed, and protective measures against harmful acoustic factors are presented. In addition, the impact of electromagnetic pulses on the components of the DDSP-WORLD vocoder is investigated, along with the regulatory framework and methods for improving system robustness against impulsive disturbances.

ЗМІСТ

ВСТУП

Актуальність теми.

Синтез співочого мовлення є однією з найбільш складних і водночас перспективних задач сучасної обробки мовлення та аудіосигналів. На відміну від класичного синтезу мовлення (Text-to-Speech), синтез співочого голосу вимагає високоточного відтворення висоти основного тону, ритмічної структури, тривалостей нот, а також експресивних характеристик виконання, таких як вібрато, легато та динамічні зміни гучності. Навіть незначні похибки у частоті основного тону або часовому вирівнюванні призводять до помітних артефактів та зниження перцептивної якості синтезу.

Сучасні методи синтезу співочого мовлення активно використовують глибинне навчання, зокрема варіаційні, дифузійні та потокові генеративні моделі. Дифузійні підходи забезпечують високу якість синтезу, однак характеризуються значною обчислювальною складністю та високою затримкою під час генерації. Потокові методи та методи узгодження потоку дозволяють суттєво скоротити кількість ітерацій під час синтезу, проте потребують ефективного представлення даних та стабільних умовлювальних механізмів.

Окремою проблемою є високі вимоги до обчислювальних ресурсів і обсягу навчальних даних, що обмежує практичне застосування багатьох сучасних моделей. У цьому контексті актуальним є поєднання генеративного моделювання в компактному латентному просторі з диференційовною цифровою обробкою сигналів, зокрема вокодерами сімейства DDSP-WORLD, які дозволяють явно моделювати гармонічну та шумову складові аудіосигналу.

Таким чином, розробка ефективного методу синтезу співочого голосу на основі узгодження потоку в латентному просторі у поєднанні з DDSP-WORLD вокодером є актуальним науково-практичним завданням, що відповідає сучасним тенденціям розвитку штучного інтелекту та обробки аудіосигналів.

Мета і задачі дослідження.

Метою кваліфікаційної роботи є розробка та дослідження методу синтезу співочого голосу на основі узгодження потоку в латентному просторі з використанням DDSP-WORLD вокодера, який забезпечує високу перцептивну якість синтезу за знижених обчислювальних витрат.

Для досягнення поставленої мети в роботі необхідно розв’язати такі задачі:

– проаналізувати сучасні підходи до синтезу мовлення та синтезу співочого мовлення;

– дослідити методи латентного подання аудіосигналів та залишкового векторного квантування;

– розробити архітектуру генеративної моделі на основі умовного узгодження потоку;

– спроєктувати та реалізувати аудіо-автокодер з DDSP-WORLD вокодером;

– розробити модель ознак музичної партитури, що враховує фонемні, нотні та часові характеристики;

– реалізувати процес навчання та оптимізації моделі;

– провести експериментальне та абляційне оцінювання якості синтезу;

– виконати аналіз і візуалізацію роботи окремих компонентів моделі.

Об’єкт дослідження.

Об’єктом дослідження є процес синтезу співочого мовлення на основі генеративних моделей глибинного навчання та методів диференційовної цифрової обробки аудіосигналів.

Предмет дослідження.

Предметом дослідження є методи та моделі узгодження потоку в латентному просторі аудіосигналів, архітектури генеративних нейронних мереж та способи їх поєднання з DDSP-WORLD вокодером для синтезу співочого голосу.

Методи дослідження.

У роботі використано такі методи дослідження:

– методи глибинного навчання та нейронних мереж;

– генеративні моделі на основі узгодження потоку;

– методи латентного моделювання та залишкового векторного квантування;

– методи диференційовної цифрової обробки сигналів;

– чисельні методи інтегрування диференціальних рівнянь;

– об’єктивні та суб’єктивні методи оцінювання якості синтезу мовлення.

Наукова новизна одержаних результатів.

Наукова новизна кваліфікаційної роботи полягає в такому:

– уперше запропоновано метод синтезу співочого голосу, що поєднує узгодження потоку в латентному просторі з DDSP-WORLD вокодером, що дало змогу підвищити продуктивність синтезу співочого мовлення у порівнянні з дифузійними моделями;

– розроблено архітектуру генеративної моделі з умовним латентним моделюванням, адаптовану до задачі синтезу співочого мовлення, що дало на практиці реалізувати запропонований метод;

– удосконаленно підхід генерації співочого мовлення за рахунок запропонованого методу, який забезпечує порівняльний рівень перцептивної якості синтезу відносно дифузійних моделей за суттєво меншої кількості ітерацій генерації, що приводить до зменшення часу синтезу.

Практичне значення одержаних результатів.

Практичне значення роботи полягає у можливості використання розробленого методу синтезу співочого голосу в програмних застосунках генерації вокалу, музичних редакторах та дослідницьких системах обробки аудіосигналів. Запропонований підхід дозволяє зменшити обчислювальні витрати під час синтезу без істотної втрати перцептивної якості, що робить його придатним для використання в умовах обмежених ресурсів. Окремі компоненти розробленої архітектури можуть бути використані як складові інших систем синтезу мовлення або аудіогенерації та слугувати основою для подальших наукових досліджень.

Публікації.

Результати кваліфікаційної роботи апробовані на науковому воркшопі Applied Information Technologies and Artificial Intelligence Systems (AITAIS), який відбувся 18–19 грудня 2025 року, а також на XIII науково-технічній конференції Тернопільського національного технічного університету імені Івана Пулюя «Інформаційні моделі, системи та технології» (17–18 грудня 2025 року) у вигляді тез доповідей.

1. S. Lupenko, M. Klishch, V. Yatsyshyn, O. Pastukh, “Singing voice synthesis via latent flow matching and differentiable digital signal processing,” in Proc. Applied Information Technologies and Artificial Intelligence Systems (AITAIS), Ternopil, Ukraine, Dec. 18–19, 2025.

2. М. Кліщ, В. Яцишин, “Структурні обмеження для підвищення ефективності генеративного моделювання аудіосигналів,” in Proc. XIII наук.-техн. конф. “Інформаційні моделі, системи та технології,” Тернопіль, Україна, 17–18 грудня 2025 р.

Структура роботи.

Кваліфікаційна робота включає пояснювальну записку та графічну частину. Пояснювальна записка складається зі вступу, чотирьох розділів, загальних висновків, списку використаних джерел і додатків. Обсяг пояснювальної записки становить 73 аркуші формату А4. Графічна частина виконана у вигляді презентаційних матеріалів та налічує 17 слайдів.

РОЗДІЛ 1 ОГЛЯД СУЧАСНИХ МЕТОДІВ СИНТЕЗУ СПІВОЧОГО МОВЛЕННЯ

1.1 Задачі синтезу мовлення та синтезу співочого мовлення

Синтез співочого голосу (Singing Voice Synthesis, SVS) спрямований на генерацію виразного співу на основі музичної партитури та тексту пісні. На відміну від синтезу мовлення (Text-to-Speech, TTS), задачі SVS вимагають точного відтворення висоти тону, ритму та виконавських нюансів, таких як вібрато й легато.

Попри концептуальну схожість, SVS ставить перед дослідниками значно складніші виклики, ніж TTS. По-перше, контур висоти тону у співі охоплює значно ширший динамічний діапазон і потребує точності на рівні кадрів: навіть незначні відхилення частоти основного тону F_0 можуть призводити до перцептивно неприродного або дисонантного звучання. По-друге, часова структура співу визначається музичним ритмом і тривалістю нот, а не природною просодією мовлення. Це створює сильну залежність між вирівнюванням фонем і музичним таймінгом, за якої некоректні тривалості або зсуви початку звуків істотно спотворюють розбірливість тексту. По-третє, такі характеристики виконавської виразності, як вібрато, портаменто та динамічні зміни гучності, є критично важливими для природності співу, але залишаються складними для моделювання у стандартних архітектурах синтезу мовлення. Крім того, співочі датасети зазвичай є меншими та менш різноманітними порівняно з корпусами мовлення, що обмежує стійкість суто даноорієнтованих підходів.

Протягом останніх двох десятиліть синтез співочого голосу еволюціонував від конкатенативних методів до нейронних генеративних парадигм. Ранні системи, такі як Vocaloid [1] та UTAU, ґрунтувалися на конкатенативному відтворенні записаних фонем, забезпечуючи ручний контроль, але обмежену виразність. Статистичні моделі, зокрема Sinsy [2], запровадили використання прихованих марковських моделей для передбачення

тривалостей нот і висоти тону, що дозволило автоматизувати процес ціною надмірно згладженого тембру. Із розвитком глибинного навчання такі архітектури, як XiaoiceSing [3] та DeepSinger [4], почали використовувати трансформерні енкодери для неавторегресивного передбачення, суттєво покращивши стабільність висоти тону та ритму.

Сучасні нейронні системи SVS вирішують ці проблеми за допомогою варіаційного, дифузійного або потокового генеративного моделювання. Попри те, що такі архітектури, як VISinger [5], досягають високої якості та здатні узагальнювати за обмеженої кількості даних, вони часто страждають від невідповідності між розподілами апостеріорного (аудіо) та апріорного (музична партитура) просторів під час навчання й інференсу, що може призводити до неточного відтворення висоти тону та помилок у вимові [6]. Дифузійні моделі, зокрема DiffSinger [7], виконують генерацію безпосередньо в акустичному просторі та забезпечують високу якість синтезу, проте потребують багатокрокового ітеративного денойзингу, що зумовлює повільний інференс і значні обчислювальні витрати. Система HiddenSinger [6] здійснює дифузійне моделювання у компактному латентному просторі, сформованому нейронним аудіокодеком, що дозволяє суттєво зменшити розмірність задачі. Проте, навіть у латентному просторі дифузійні підходи залишаються стохастичними та потребують десятків або сотень кроків денойзингу, що обмежує швидкість синтезу.

1.1.1 Моделі синтезу співочого голосу

Сучасні дослідження у сфері синтезу співочого голосу (SVS) активно розвиваються завдяки поєднанню методів глибинного навчання, нейронних аудіокодеків, дифузійних процесів, потокових (flow-based) підходів та методів цифрової обробки сигналів (DSP). Еволюція моделей SVS відображає перехід від детермінованого акустичного моделювання до імовірнісних і латентних генеративних парадигм.

Однією з представників цього покоління є система XiaoiceSing [3], яка

використовує архітектуру, подібну до FastSpeech, і поєднує фонемні, позиційні та музичні ознаки для передбачення тривалостей фонем і частоти основного тону F_0. Модель здійснює неавторегресивне акустичне передбачення, водночас враховуючи ритмічне та мелодичне умовлювання.

Серія моделей VISinger [5, 8] побудована на основі архітектури VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech). Вона інтегрує варіаційне латентне моделювання, потокові перетворення та адверсаріальне навчання для спільного моделювання акустичних ознак і реконструкції хвильової форми. VISinger 2 розширює цей підхід шляхом впровадження диференційовних блоків цифрової обробки сигналів (DDSP) та вокодера HiFi-GAN, явно розділяючи гармонійну та шумову складові сигналу. Система функціонує в режимі end-to-end, синтезуючи аудіо з частотою дискретизації 44,1 кГц, і була навчена протягом 500 тисяч кроків на датасеті тривалістю 5 годин.

Модель DiffSinger [7] застосовує імовірнісне дифузійне моделювання для генерації мел-спектрограм. Вона перетворює гаусівський шум у цільові спектрограми за допомогою умовного дифузійного процесу, а використання спрощеного («мілкого») дифузійного механізму дозволяє зменшити кількість кроків денойзингу та підвищити ефективність генерації.

HiddenSinger [6] поєднує нейронний аудіокодек із латентною дифузійною моделлю. Система кодує співочий аудіосигнал у низьковимірний латентний простір, виконує генерацію безпосередньо в цьому просторі та відновлює повносмуговий сигнал за допомогою нейронного декодера. Модель була навчена на 150 годинах даних протягом 3 мільйонів кроків із розміром пакета 32. Варіант HiddenSinger-U підтримує напівкероване навчання на непарних співочих даних.

Проблеми багатомовності та zero-shot адаптації досліджуються в TCSinger 2 [9], який забезпечує крослінгвальний і крос-виконавський синтез без додаткового донавчання. Водночас TechSinger [10] використовує парадигму узгодження потоку (flow matching) для керованого синтезу співу, надаючи

можливість явного контролю вокальних технік, таких як вібрато та придих, у багатомовному середовищі.

Загалом сучасні системи SVS охоплюють широкий спектр архітектур – від неавторегресивних моделей (XiaoiceSing) і варіаційних фреймворків на основі VITS (VISinger, VISinger 2) до дифузійних (DiffSinger), латентно-дифузійних (HiddenSinger) та потокових (TechSinger) підходів, – демонструючи поступальний розвиток у напрямку багатомовності, керованості та підвищення виразності співочого синтезу.

1.1.2 Сучасні TTS-моделі на основі узгодження потоку

Останні досягнення у моделюванні синтезу мовлення (Text-to-Speech, TTS) демонструють послідовний перехід від авторегресивних архітектур до потокових і латентних генеративних підходів, спрямованих на підвищення ефективності синтезу, узгодженості просодії та керованості процесу генерації. Низка робіт досліджує різні формулювання узгодження потоку (flow matching) і дифузійних процесів для генерації мовлення.

Модель MetaTTS [11] інтегрує стратегії попереднього навчання у TTS та досліджує, яким чином масштабне узгодження потоку може покращити узагальнювальну здатність і природність синтезу. Це дослідження є однією з перших спроб застосувати flow matching у масштабі фундаментальних мовленнєвих моделей, поєднуючи попередньо навчені акустичні представлення з генеративним моделюванням.

Продовжуючи цей напрям, F5-TTS [12] зосереджується на задачі довготривалого синтезу мовлення, роблячи акцент на вирівнюванні тексту й мовлення та часовій узгодженості. Використовуючи фреймворк узгодження потоку, модель зберігає цілісну просодію в межах протяжних висловлювань, що підкреслює придатність потокових методів для наративного й експресивного синтезу мовлення.

У роботі Matcha-TTS [13] запропоновано спрощену архітектуру умовного узгодження потоку, оптимізовану для низької затримки під час синтезу. Модель

суттєво скорочує час інференсу, зберігаючи конкурентну якість аудіо, що вказує на її потенціал для застосувань у режимі реального часу.

Комплементарний напрям досліджується в LatentSpeech [14], де для TTS застосовується латентне дифузійне моделювання. Генерація мовлення відбувається в компактному латентному просторі, а не безпосередньо у спектральному чи часовому домені, що дозволяє зменшити обчислювальні витрати без втрати акустичної деталізації. Цей підхід демонструє ефективність, яку можна досягти завдяки дифузії в латентному просторі.

Модель VoiceFlow [15] розширює потокову парадигму, використовуючи узгодження випрямленого потоку (rectified flow matching), що переосмислює динаміку процесу з метою мінімізації кількості кроків інтегрування та вартості інференсу. Отримана модель досягає порівнюваної перцептивної якості за суттєво кращої обчислювальної ефективності.

ProsodyFlow [16] поєднує умовне узгодження потоку з моделюванням просодії, інформованим великими мовленнєвими мовними моделями. Таке поєднання забезпечує явний контроль над високорівневими просодичними характеристиками, зокрема інтонацією, ритмом і наголосом, демонструючи, як лінгвістичне умовлювання може підвищити виразність нейронного синтезу мовлення.

1.2 Огляд існуючих рішень та літератури щодо вокодерів

У сучасній обробці цифрових сигналів та синтезі мовлення вокодери відіграють ключову роль як інструменти для аналізу та синтезу людського голосу. Традиційні вокодери, такі як Griffin-Lim [17], WORLD [18] і STRAIGHT [19], базуються на параметричних моделях, які аналізують та синтезують мовлення.

Зі стрімким розвитком глибокого навчання значний прогрес у синтезі мовлення досягнуто завдяки нейромережевим вокодерам. На відміну від традиційних параметричних методів, які використовують спектральні перетворення та моделі з фіксованими правилами, нейромережеві підходи

здатні навчатися безпосередньо на великих наборах аудіоданих та більш краще узагальнувати латентні параметри голосу, що дозволяє їм краще відтворювати природність голосу, інтонацію та тембр мовлення.

Серед найбільш популярних нейромережевих вокодерів можна виділити кілька основних типів:

1. Генеративні змагальні мережі – використовують дві нейромережі: генератор і дискримінатор, які навчаються разом, щоб покращувати якість синтезованого мовлення. Цей підхід використано у HiFi-GAN [20], uSFGAN [21], SiFi-GAN [22]. Одним із найбільших викликів є проблема нестабільності в процесі навчання.

2. Автоагресивні моделі – генерують аудіо послідовно, покроково прогнозуючи наступний аудіосемпл на основі попередніх. Цей підхід дає високу якість синтезу, але є обчислювально затратним. Реалізований у моделях WaveRNN [23], Wavenet [24]. Недоліками підходу є необхідність потужних обчислювальних ресурсів для обробки кожного кроку та значна кількість часу для генерації.

3. Flow-based моделі – використовують зворотні авторегресивні потоки. Прикладами є Parallel Wavenet [25], WaveGlow [26] та ClariNet [27]. Хоча ці моделі можуть швидко генерувати високоякісне мовлення, вони можуть бути менш гнучкими в порівнянні з автоагресивними моделями, особливо коли йдеться про відтворення складних мовних характеристик. Також ці моделі є складнішими у навчанні та потребують моделі-вчителя.

4. Гібридні моделі – поєднують переваги різних підходів для досягнення оптимального балансу між якістю мовлення та швидкістю обчислень. Прикладом є LPCNet [28], який використовує лінійне прогнозування (Linear Predictive Coding, LPC) разом із нейромережевими підходами, що робить його ефективним для мобільних пристроїв. Проте, поєднання різних методів може призвести до складних і важких для навчання моделей. Вони можуть вимагати більше часу та ресурсів для налаштування та оптимізації.

У роботі [29] автори представляють фреймворк, розроблений на основі

TensorFlow, який інтегрує традиційні методи цифрової обробки сигналів (DSP) із глибоким навчанням для синтезу звуку. Основна мета полягає в тому, щоб зробити класичні елементи обробки сигналів диференційованими, що дозволяє навчати їх за допомогою сучасних нейронних мереж, зберігаючи при цьому ефективність та інтерпретованість.

Таким чином, диференційована цифрова обробка сигналів (англ. Differentiable Digital Signal Processing, DDSP) є гібридним підходом, який поєднує диференційовані компоненти DSP з методами глибокого навчання, що розширює можливості обробки та генерації аудіосигналів.

Основними перевагами є:

1. Відсутність необхідності у великих авторегресивних моделях та складних функція втрат.

2. Використання інтерпретованих та зрозумілих компонентів цифрової обробки сигналів (осцилятори, фільтри, реверберація, огинаючі, адитивний синтез).

3. Забезпечується незалежне керування висотою, тембром і гучністю, дозволяючи інтуїтивно маніпулювати створеними звуками.

4. Підтримується синтез у реальному часі завдяки меншому розміру моделі.

У статті [30] представлено диференційовану версію вокодера WORLD. Запропонована модель дозволяє наскрізне навчання передачі аудіостилю, включаючи такі завдання, як перетворення голосу та передача тембру. У цій статті доводиться, що WORLD зберігає переваги, такі як збереження висоти тону та чітке виділення акустичних характеристик, яких бракує сучасним нейронним вокодерам. На відміну від оригінального вокодера WORLD, ця версія дозволяє навчатися на основі градієнта, що робить її придатною для фреймворків глибокого навчання.

Щоб зменшити кількість параметрів, спектральну огинаючу стискають у формат log Mel-спектрограми, що робить її більш ефективною для нейронних мереж. Також робиться стиснення розтиснення аперіодичної складової за

допомогою інтерполяції значень.

Також автори [30] пропонують альтернативний метод трансферу голосу на основі витягування сигналу збудження безпосередньо з вихідного аудіо. Цей метод не потребує синтезу, тому потенційно є швидшим та як зазначають автори можливо більш реалістичним для задачі SVC.

У [31] представлено Glottal-flow LPC Filter (GOLF), метод синтезу співочого голосу (SVS) на основі DDSP. На відміну від традиційних вокодерів на основі глибокого навчання, які покладаються на вхідну мелоспектрограму, GOLF використовує глоткову модель та рекурсивний фільтр для кращої імітації динаміки голосового тракту людини.

У [32] представлено SawSing – вокодер на основі DDSP, який синтезує співочі голоси з Mel-спектрограм. SawSing розроблено для усунення фазових розривів – поширеної проблеми в існуючих нейронних вокодерах, яка призводить до неприродних артефактів, особливо у довгих співочих висловлюваннях.

У [33] представлено надлегкий DDSP-вокодер. У роботі описано вокодера аналонічного до DDSP WORLD з системою TTS з метою зменшення використання обчислювальних ресурсів.

У [34] подано огляд та підсумовано значну кількість робіт по диференційованій обробці сигналів. Також автори описали значну кількість можливих застосувань підходу. У роботі були представленні відкриті питання щодо диференційованої обробки сигналів.

Огляд демонструє, що підхід DDSP є перспективним напрямом у синтезі звуку, який поєднує класичні DSP-методи з можливостями глибокого навчання. Основною перевагою є збереження інтерпретованості та ефективності традиційних методів обробки сигналів при одночасному використанні градієнтного навчання.

1.3 Методи теселяції простору та залишкове векторне квантування

Обґрунтування необхідності теселяції простору полягає в тому, що латентні представлення, які формуються генеративною моделлю, зазвичай характеризуються значною варіативністю. Така розбіжність у значеннях латентного простору ускладнює завдання декодера, оскільки він має відтворювати сигнал з надмірно широкого та нерівномірно розподіленого простору. Теселяція дозволяє розділити латентний простір на підобласті, що зменшує локальну варіативність та уніфікує розподіл ознак.

1.3.1 Векторна квантизація

Векторна квантизація використується для компактного подання багатовимірних даних за допомогою обмеженої кількості кодових векторів (кодбуку) [35]. У контексті даної роботи цей метод розглядається для зменшення варіативності латентного простору.

Формально процес векторної квантизації з відомим кодбуком $C = \{c_1, c_2, ..., c_K\}$ для деякого сигналу $z \in \mathbb{R}^d$ визначається як пошук $\hat{k}$:

$$\hat{k} = \arg\min_k ||z - c_k||^2. \tag{1.1}$$

З точки зору геометрії, векторна квантизація визначає розбиття простору з деякою заданою метрикою (у даному випадку Евклідовою відстанню) на області Вороного. Кожен вектор z потрапляє до тієї комірки, центр якої є найближчим кодовим вектором.

Введемо квантизуючу функцію $q(\cdot, C) : \mathbb{R}^d \to C$, яка кожному вхідному вектору z ставить у відповідність найближчий кодовий вектор з C:

$$q(z, C) = \arg\min_{c_k \in C} ||z - c_k||^2. \tag{1.2}$$

У випадках, коли з контексту зрозумілий кодбук C, функцію позначатимемо як $q(z)$.

Після визначення процесу векторної квантизації природно постає питання: яким чином формувати кодбук C. Мета формування кодбука C полягає в тому, щоб його елементи відображали статистичну структуру даних. Якщо ж кодові вектори вибирати випадково або розташовувати на регулярній сітці, це призводить або до значних спотворень, або до експоненційного зростання кількості кодів у високих розмірностях.

Отже, необхідно знайти такий кодбук $\hat{C}$, що дозволяє мінімузувати середньо квадратичну різницю між даними та їх квантовими представленнями:

$$\hat{C} = \arg\min_{C} \mathbb{E}_{z\sim p(z)} \|z - q(z, C)\|^2 \tag{1.3}$$

Одним з методів є застосування алгоритму k-means. Таким чином задача пошуку кодбуку зводиться до кластеризації даних. Проте, з метою вбудувати векторну квантизацію у архітектуру нейронної мережі розглянемо VQ-VAE.

1.3.2 VQ-VAE

VQ-VAE [36] – архітектура автоенкодера, яка поєднує векторну квантизацію з варіантивним автоенкодером. Модель складається з трьох блоків:

1. Енкодера, який на основі вхідних даних x, формує латентний вектор z.
2. Векторного квантизатора $q(z)$, який формує квантизоване представлення.
3. Декодера, що на основі $q(z)$ передбачає x.

Автори моделі пропонують доповнити реконструкційну функцію втрат двома додатковими функціями: функцією втрат кодбука та комітмент-функцією втрат.

Функція втрат кодбуку забезпечує відповідність кодових векторів статистичній структурі латентного простору z:

$$\mathcal{L}_{codebook} = \|sg[z] - q(z)\|^2, \tag{1.4}$$

де $sg[\cdot]$ – оператор зупинки градієнту.

Комітмент-функція втрат забезпечує зворотнє: значення z наближуються до $q(z)$, що дозволяє уникнути ситуації, коли енкодер навчається ефективніше квантизатора. Функція визначається як:

$$\mathcal{L}_{commitment} = \beta||z - sq[q(z)]||^2, \tag{1.5}$$

де β – гіперпараметр, що визначає наскільки $q(z)$ буде «притягувати» z.

Метод запропонований у VQ-VAE набує розвитку в роботах [37] та [38].

1.3.3 Залишкова векторна квантизація

Для зменшення варіативності латентного простору пропонується застосувати залишкову векторну квантизацію (ЗВК), аналогічно як це було зроблено у роботі [6].

Власне метод було запроповано у роботі [39].

ЗВК апроксимує сигнал $z \in \mathbb{R}^d$ як суму L квантизованих векторів, кожен з яких обраних з кодбуку $C^{(i)}, i = 1, 2, ..., L$ фіксованого розміру M_i:

$$q(z) = \sum_{i=1}^{L} q_i(r_i), \tag{1.6}$$

де $q_i : \mathbb{R}^d \to C^{(i)}$ – квантизуюча функція для кодбуку $C^{(i)}$, r_i – залишковий вектор на i-ому кроці квантизації:

$$r_1 = z, r_i = z - \sum_{j=1}^{i-1} q_j(r_j). \tag{1.7}$$

Таким чином кожен залишковий вектор r_i буде доповнювати різницю між сигналом z та його квантизованим варіантом.

1.4 Актуальність завдання

У контексті синтезу співочого мовлення спостерігається тенденція до збільшення якості генерація, що супроводжується збільшенням обсягу датасету

та часу тренування. Проте, в умовах малоресурсної або середньоресурсної мови, коли обсяг доступних датасетів є суттєво обмеженим, виникає необхідність у методах, що будуть ефективно використовувати дані.

Іншим викликом є інфраструктура для тренування якісних моделей, що часто потребує тривалого тренування із залученням значної кількості обчислювальних ресурсів. Це формує запит на методи, які будуть менш обчислювально затратними.

1.5 Висновки

У аналітичній частині роботи досліджено сучасний стан задач синтезу мовлення та синтезу співочого мовлення. Проаналізовано основні класи моделей SVS, зокрема неавторегресивні, варіаційні, дифузійні та потокові підходи, а також сучасні TTS-моделі на основі узгодження потоку.

Окрему увагу приділено вокодерам та методам параметризованого синтезу, зокрема підходам, що поєднують нейронні моделі з цифровою обробкою сигналів. Розглянуто методи дискретизації та теселяції латентного простору, включно з векторним та залишковим векторним квантування.

На основі проведеного аналізу встановлено, що дифузійні моделі забезпечують високу якість синтезу, однак мають значну обчислювальну складність, тоді як потокові моделі демонструють кращий баланс між якістю та швидкодією. Це обґрунтовує доцільність дослідження підходів узгодження потоку в латентному просторі у поєднанні з диференційовними вокодерами.

РОЗДІЛ 2 МЕТОДИ ТА АРХІТЕКТУРА МОДЕЛІ СИНТЕЗУ СПІВОЧОГО ГОЛОСУ НА ОСНОВІ УЗГОДЖЕННЯ ПОТОКУ

У цьому розділі представлено запропоновану архітектуру синтезу співочого голосу з умовним латентним моделюванням, яка поєднує латентний предиктор на основі потокових моделей із автокодером на базі диференційованої цифрової обробки сигналів (DDSP) для реконструкції аудіосигналу, як показано на рис. 2.1.

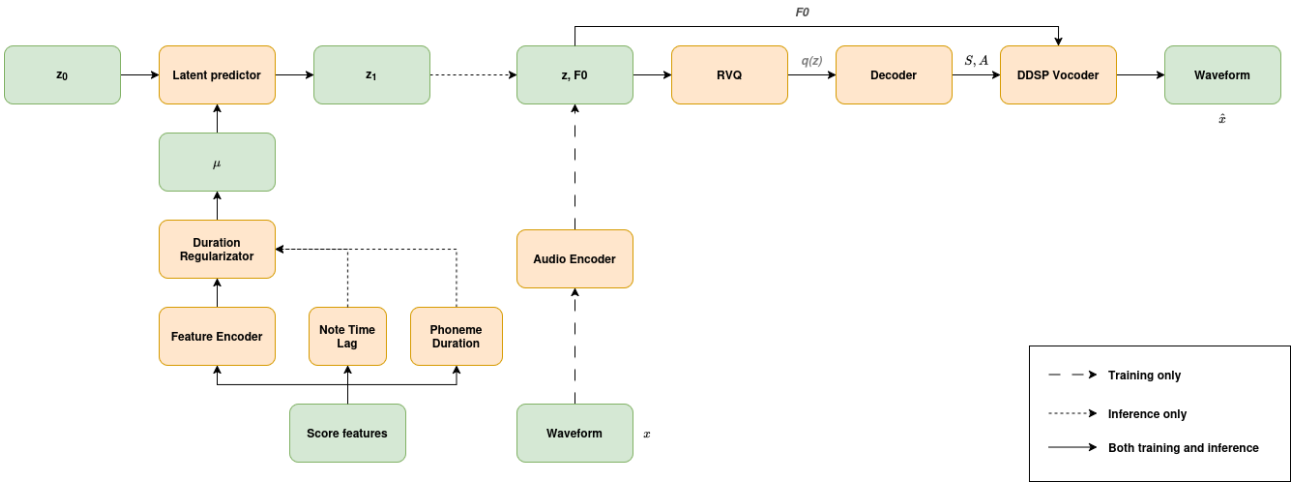

Рисунок 2.1 – Огляд пропонованої архітектури

Загальна архітектура моделі складається з кількох взаємопов'язаних функціональних модулів, кожен з яких відповідає за окремий етап процесу синтезу. Умовні ознаки, отримані з музичної партитури та лінгвістичної інформації, використовуються для керування генеративною динамікою в латентному просторі, тоді як аудіо-автокодер забезпечує відображення латентних представлень у часовий аудіосигнал. Центральним елементом запропонованої системи є модуль узгодження потоку, який відповідає за передбачення латентної траєкторії між апріорним та цільовим розподілами. Подальші підрозділи цього розділу детально описують теоретичні засади узгодження потоку, архітектуру генеративної моделі та допоміжні компоненти системи.

2.1 Узгодження потоку векторного поля

Узгодження потоку (англ. Flow Matching, FM) [40] – це підхід до навчання безпосереднього відображення з простого розподілу (наприклад, гаусового) у складний цільовий розподіл даних за допомогою векторного поля. На відміну від дифузійних моделей, де процес задається через стохастичні диференційні рівняння, у Flow Matching розглядається детерміноване диференціальне рівняння, яке описує безперервну траєкторію перетворення розподілу.

Пропонується використати цей метод для передбачення вбудування z, моделюючи векторний поле $v_t(z)$, який визначає перетворення розподілу $p_0 = \mathcal{N}(0, 1)$ у розподіл p_1, що апроксимує розподіл даних латентного простору:

$$\frac{d\psi_t(z)}{dt} = v_t(\psi_t(z)) \tag{2.1}$$

де $\psi : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ – часозалежне векторний потік:

$$\psi_t(z) = (1 - t)z + tz_1. \tag{2.2}$$

Функція втрат визначається наступним чином:

$$\mathcal{L}_{fm} = \mathbb{E}_{t\sim U[0,1], z_0\sim p_0, z_1\sim p_1} ||(z_1 - z_0) - v_t(\psi_t(z_0))||^2 \tag{2.3}$$

2.1.1 Модифікована функція втрат

Функція втрат (2.3) має низку обмежень, які знижують ефективність її використання на практиці. Зокрема, основними проблемами є:

1. Відсутність можливості комбінування з іншими функціями втрат (реконструктивними, регуляризаційними), що обмежує гнучкість підходу.

2. Неоптимальна поведінка при малих значення t: прогнозування траєкторії поблизу $t = 0$ є складнішим і менш стабільним, ніж для значень поблизу $t = 1$. Проте, від якості ранніх результатів генерації буде залежати якість кінцевого сформованого об’єкта.

Як наслідок, навіть після навчання не гарантовано, що отримані потоки будуть оптимальними з точки зору довжини траєкторії та узгодженості з розподілом.

Як показано у [41] функція втрат (2.3) еквівалентна до:

$$\mathcal{L}_{fm} = \mathbb{E}_{t\sim U[0,1],z_0\sim p_0,z_1\sim p_1}\left(\frac{1}{t^2}\|z_1 - z_\theta(\psi_t(z_0))\|^2\right), \tag{2.4}$$

де z_θ – передбачення моделі.

У роботі [42] пропонується інший варіант функції втрат:

$$\mathcal{L}_{fm} = \mathbb{E}_{t\sim U[0,1],z_0\sim p_0,z_1\sim p_1}\left(\frac{1}{1-t}\|z_1 - z_\theta(\psi_t(z_0))\|^2\right). \tag{2.5}$$

Загалом обидві функції (2.4) та (2.5) дозволяють використати передбачення моделі як передбачення цільового об’єкта. Отже, до θ може бути застосована реконструктивна функція втрат.

Як показано у [41, 43] функція втрат може бути узагальненою до:

$$\mathcal{L}_{fm} = \mathbb{E}_{t\sim T,z_0\sim p_0,z_1\sim p_1}\left(w(t)\|z_1 - z_\theta(\psi_t(z_0))\|^2\right), \tag{2.6}$$

де T – функція розподілу значень часової змінної в інтервалі (0; 1).

Цей варіант функції втрат пропонується включити до пропованого рішення, оскільки це дозволить застосувати реконструктивні функції втрат та навчити модель типу End-to-End.

2.1.2 Умовне узгодження потоку

У межах даної роботи узгодження потоку застосовується для передбачення латентних вбудувань z в умовному режимі. Для цього моделюється часово-залежне векторне поле $v_t(z \mid \mu)$, де μ позначає умовні ознаки, отримані з енкодера, що описують вхідні акустичні або музичні характеристики. Таке формулювання дозволяє поєднати механізм узгодження потоку з умовною генерацією латентних представлень.

Відповідне диференціальне рівняння набуває вигляду:

$$\frac{d\psi_t(z \mid \mu)}{dt} = v_t(\psi_t(z \mid \mu)), \tag{2.7}$$

де ψ_t – часозалежний потік у латентному просторі. Як і в базовому випадку, траєкторія перетворення задається лінійною інтерполяцією між вибіркою з апріорного розподілу $z_0 \sim p_0$ та цільовим латентним представленням $z_1 \sim p_1$:

$$\psi_t(z_0) = (1 - t)z_0 + tz_1. \tag{2.8}$$

Навчання умовної моделі узгодження потоку здійснюється з використанням репараметризованої функції втрат на основі оптимального транспорту, запропонованої у [42]:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t\sim U[0,1],\, z_0\sim p_0,\, z_1\sim p_1}\left(\frac{1}{1-t}\left\|z_1 - z_\theta(\psi_t(z_0) \mid \mu)\right\|_2^2\right), \tag{2.9}$$

де z_θ – передбачення моделі. Ця форма функції втрат є окремим випадком узагальненої формули (2.6) з відповідною ваговою функцією $w(t) = \frac{1}{1-t}$.

2.1.3 Узгодження умовних ознак

Для покращення зв’язку між умовними ознаками μ та цільовим латентним представленням z_1 додатково вводиться допоміжна реконструктивна функція втрат:

$$\mathcal{L}_{\text{feature}} = \mathbb{E}\|\mu - z_1\|_2^2. \tag{2.10}$$

Запровадження цієї складової стимулює енкодер формувати умовні ознаки, які містять прогностичну інформацію про цільове латентне вбудування, не порушуючи при цьому динаміку узгодження потоку. Таким чином, загальна функція втрат може бути розширена реконструктивними компонентами, що дозволяє навчати модель у режимі End-to-End.

2.1.4 Архітектура моделі

Запропонована архітектура предиктора FlowSinger базується на принципах, закладених у моделі Matcha-TTS [13], та наслідує U-Net-подібну структуру з ієрархічною багаторівневою обробкою часових ознак. Такий підхід дозволяє ефективно поєднувати локальні та глобальні часові залежності у задачі оцінювання векторного поля для узгодження потоку в латентному просторі. Загальну схему архітектури наведено на рис. 2.2.

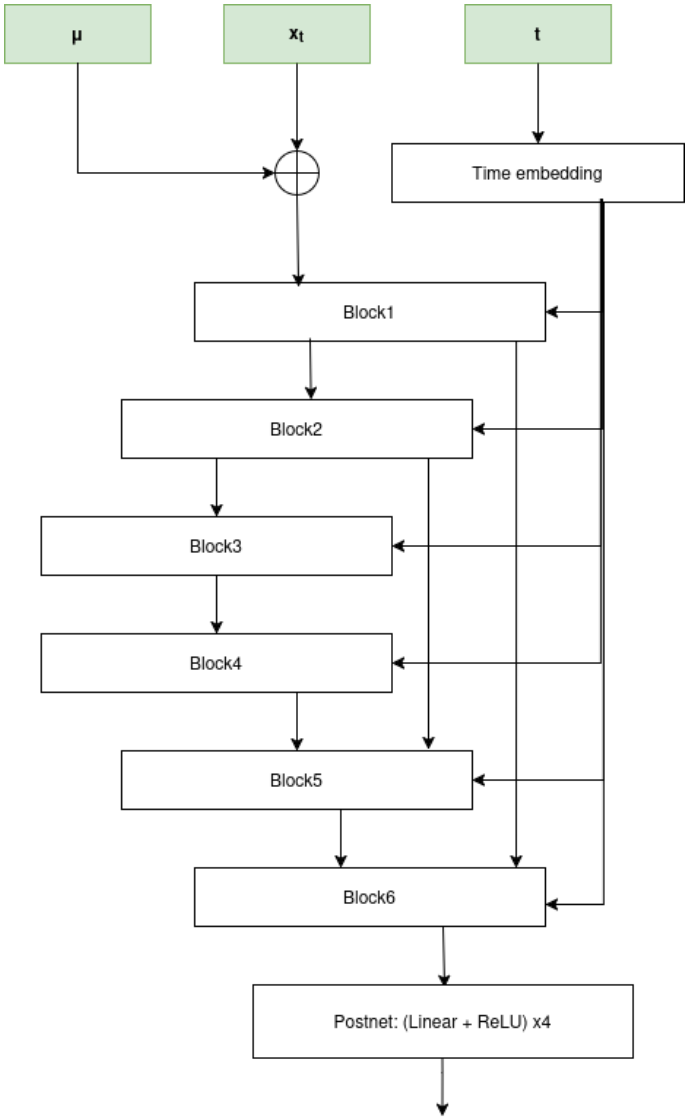

Рисунок 2.2 – Схема архітектури предиктора векторного поля v_t

Модель організована у вигляді послідовності блоків, які формують низхідну, проміжну та висхідну частини мережі. Низхідна гілка відповідає за поступове агрегування локального контексту та зменшення розмірності

ознак, тоді як висхідна частина забезпечує відновлення деталізованих часових структур із використанням інформації з ранніх рівнів. Подібно до архітектур типу U-Net, між симетричними рівнями використовуються резидуальні з'єднання, що сприяє стабільному поширенню інформації та зменшенню втрат деталізації під час багаторівневого перетворення ознак.

Часова змінна t, яка параметризує крок у задачі узгодження потоку, проєктується у компактне часове вкладення та явно додається до вхідних ознак на кожному рівні мережі. Таке умовлювання дозволяє моделі коректно репрезентувати безперервну динаміку векторного поля та адаптувати обчислення до поточного етапу інтегрування.

Кожен рівень архітектури реалізовано у вигляді окремого блоку Flow Matching, схему якого наведено на рис. 2.3. Блок поєднує одномірний згортковий шар, що відповідає за моделювання локального часовго контексту, та модуль Emformer [44], який забезпечує ефективне врахування довготривалих часових залежностей за обмежених обчислювальних витрат.

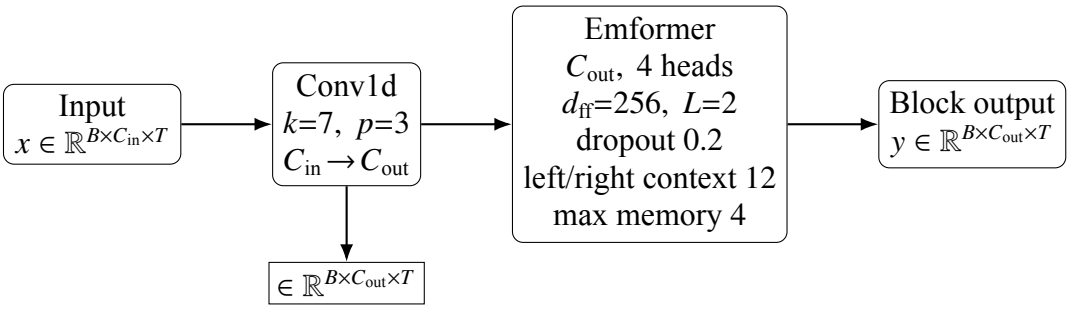

Рисунок 2.3 – Схема блоку предиктора

У базовій архітектурі Matcha-TTS [13] як основні часові блоки використовуються трансформерні шари, запозичені з BigVGAN [45], які застосовуються для моделювання часових залежностей у послідовностях акустичних ознак.

У запропонованій моделі трансформерні блоки BigVGAN замінено на модулі Emformer [44], які поєднують локальну самоувагу з механізмом обмеженої пам'яті для врахування довготривалого контексту. Це дозволяє

зменшити обчислювальні витрати, знизити затримку інференсу та забезпечити можливість потокового синтезу у задачі моделювання латентної динаміки.

2.2 Аудіо-автокодер

У запропонованій архітектурі використовується автокодер на основі диференційованої цифрової обробки сигналів (Differentiable DSP). Дотримуючись підходу HiddenSinger [6], у вузьке місце (bottleneck) інтегровано блоки залишкового векторного квантування (Residual Vector Quantization, RVQ), що забезпечує компактне латентне представлення вхідних ознак.

2.2.1 Енкодер

Нехай $x \in \mathbb{R}^N$ – вхідний аудіосигнал, де N є кількістю аудіосемплів. Енкодер $E(\cdot)$ виділяє латентну послідовність ознак:

$$z = E(x) \in \mathbb{R}^{d \times T}, \tag{2.11}$$

де d позначає розмірність ознак, а T – кількість закодованих часових кадрів.

Енкодер перетворює сирий аудіосигнал у часово-узгоджене латентне представлення, яке узагальнює його релевантну акустичну структуру. Це представлення зберігає інформацію про спектральну форму, енергію та часову еволюцію сигналу, водночас відкидаючи дрібнодетальні семплові характеристики, необхідність яких відсутня для моделювання вищого рівня.

Дотримуючись принципів архітектури DDSP [29], у роботі використано компоненти енкодера частоти основного тону F_0 та енкодера латентних ознак z, водночас енкодер гучності не застосовується, оскільки запропонована модель не використовує ознаки гучності. У даній постановці загальна амплітуда сигналу та сприймана гучність вважаються такими, що неявно кодуються у латентній змінній z.

Енкодер F_0 базується на попередньо навченому оцінювачі висоти тону CREPE [46]. Для енкодера z використано ту саму процедуру виділення ознак та

мережеву структуру, що й у [29], яка відображає MFCC-подібні представлення у компактне латентне вкладення $z(t)$ за допомогою шару GRU.

2.2.2 Залишкове векторне квантування

Залишкове векторне квантування (RVQ) апроксимує сигнал $z \in \mathbb{R}^d$ у вигляді суми L квантизованих векторів, кожен з яких вибирається з кодбука $C^{(i)},\ i = 1, 2, \ldots, L$, фіксованого розміру M_i:

$$q(z) = \sum_{i=1}^{L} q_i(r_i), \tag{2.12}$$

де $q_i : \mathbb{R}^d \to C^{(i)}$ – функція квантування для i-го кодбука $C^{(i)}$, а r_i – залишковий вектор на i-му кроці квантування:

$$r_1 = z, \quad r_i = z - \sum_{j=1}^{i-1} q_j(r_j). \tag{2.13}$$

Така структура забезпечує високоточну реконструкцію при використанні компактних кодбуків, формує дискретний апріорний розподіл у латентному просторі та накладає структурні обмеження на латентний многовид.

Модуль залишкового векторного квантування реалізовано з використанням 8 квантизаторів, кожен з яких містить кодбук розміром 1024 вектори розмірності 128.

2.2.3 Декодер та вокодер

Декодер адаптовано з архітектури, запропонованої в [33], із використанням блоків Emformer [44]. Вокодер побудовано відповідно до принципів [30, 33], поєднуючи диференційовану синтезу сигналів із декодуванням спектральних параметрів. Він оперує трьома компонентами ознак: основною частотою F_0, спектральною огинаючою S та коефіцієнтом аперіодичності A.

Сигнал збудження для гармонічної складової визначається як сума пилкоподібних сигналів [47]:

$$e_h(t) = \sum_{k=1}^{K} \frac{1}{k} \sin(\phi_k(t)), \tag{2.14}$$

де K – кількість гармонік, а $\phi_k(t)$ – миттєва фаза k-тої гармоніки, яка визначається як:

$$\phi_k(t) = 2\pi k \int_0^t F_0(\tau)\, d\tau. \tag{2.15}$$

Далі сигнал збудження пропускається через частотний фільтр для формування гармонічної складової:

$$x_h = \mathcal{F}^{-1}((1 - A)S\mathcal{F}(e_h)), \tag{2.16}$$

де $\mathcal{F}$ та $\mathcal{F}^{-1}$ позначають пряме та обернене короткочасні перетворення Фур'є (STFT) з довжиною вікна 2048 семплів та перекриттям 75%, A – коефіцієнт аперіодичності, а S – спектральний формуючий фільтр.

Початковий сигнал збудження для шумової складової визначається як гаусівський білий шум:

$$e_n(t) \sim \mathcal{N}(0, 1). \tag{2.17}$$

Шумова складова формується шляхом застосування того самого спектрального фільтра з урахуванням аперіодичності:

$$x_n = \mathcal{F}^{-1}(AS\mathcal{F}(e_n)). \tag{2.18}$$

Синтезований співочий сигнал отримується як зважена сума гармонічної та шумової складових:

$$\hat{x} = \lambda_h x_h + \lambda_n x_n, \tag{2.19}$$

де λ_h та λ_n – коефіцієнти, що керують відносним внеском та амплітудою гармонічної й шумової складових відповідно.

Відповідно до [30] для параметрів S та A застосовується стиснення та відновлення.

Лінійна спектральна огинаюча S відображається в лог-мел домен за допомогою мел-фільтробанку M:

$$S_c = \log_{10}(MS + \varepsilon), \tag{2.20}$$

а під час синтезу реконструюється як

$$S = M_0\, 10^{S_c} - \varepsilon, \tag{2.21}$$

де $M_0 = \max(M^{-1}, 0)$ – невід'ємна псевдообернена матриця до M, яка застосовується поелементно для уникнення артефактів негативної реконструкції.

Для аперіодичності декодер передбачає стиснене представлення A_c з 16 каналами, яке лінійно інтерполюється вздовж частотної осі для отримання повновимірного 513-вимірного вектора аперіодичності A, що використовується диференційованим модулем синтезу сигналу.

2.2.4 Архітектура змагально-генеративної моделі

Дотримуючись підходу [20], у роботі використовується мультиперіодний дискримінатор (Multi-Period Discriminator, MPD) та мультишкальний дискримінатор (Multi-Scale Discriminator, MSD) для покращення перцептивної якості згенерованого аудіосигналу. MPD призначений для виявлення періодичних структур, синхронізованих з основною частотою, на кількох часових періодах, тоді як MSD аналізує сигнал на різних масштабах з метою оцінювання спектральної узгодженості та довготривалої когерентності. Обидва дискримінатори навчаються спільно з генератором із використанням адверсаріальних цільових функцій типу найменших квадратів та додаткового члена узгодження ознак (feature matching).

Генератор G та дискримінатори $\{D_{\text{MPD}}^{(i)}\}_{i=1}^{N_p}$ і $\{D_{\text{MSD}}^{(j)}\}_{j=1}^{N_s}$ оптимізуються за допомогою адверсаріальних функцій втрат типу least-squares [48]. Для еталонного сигналу x та згенерованого сигналу $\hat{x} = G(\cdot)$ цільові функції визначаються таким чином.

Функція втрат генератора має вигляд:

$$\mathcal{L}_G = \mathbb{E}_{\hat{x}}\left[\sum_{i=1}^{N_p}(D_{\text{MPD}}^{(i)}(\hat{x})-1)^2 + \sum_{j=1}^{N_s}(D_{\text{MSD}}^{(j)}(\hat{x})-1)^2\right], \tag{2.22}$$

Функція втрат дискримінаторів визначається симетрично:

$$\begin{aligned}\mathcal{L}_D = \mathbb{E}_{x,\hat{x}}\Bigg[&\sum_{i=1}^{N_p}\left((D_{\text{MPD}}^{(i)}(x)-1)^2 + (D_{\text{MPD}}^{(i)}(\hat{x}))^2\right)+\\ &+\sum_{j=1}^{N_s}\left((D_{\text{MSD}}^{(j)}(x)-1)^2 + (D_{\text{MSD}}^{(j)}(\hat{x}))^2\right)\Bigg].\end{aligned} \tag{2.23}$$

Окрім адверсаріальних складових, для стабілізації навчання та покращення перцептивної якості у роботі використовується функція узгодження ознак, запропонована у [20]. Ідея цього підходу полягає у мінімізації різниці між внутрішніми активаціями дискримінаторів для реального та згенерованого сигналів.

Нехай $D_l^{(i)}(\cdot)$ позначає вихід l-го шару i-го дискримінатора, а L_i – кількість шарів у відповідному дискримінаторі. Тоді функція узгодження ознак визначається як:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x,\hat{x}}\left[\sum_{i}\sum_{l=1}^{L_i}\left\|D_l^{(i)}(x) - D_l^{(i)}(\hat{x})\right\|_1\right], \tag{2.24}$$

де сума береться за всіма дискримінаторами мультиперіодної та мультишкальної груп.

2.3 Модель ознак музичної партитури

Модель ознак музичної партитури відповідає за перетворення дискретної інформації, що міститься у нотному записі та тексті пісні, у безперервне часово-узгоджене представлення, придатне для подальшого генеративного моделювання.

2.3.1 Енкодер фонемних та нотних ознаки

Енкодер фонемних та нотних ознак формує умовне представлення на основі фонемної послідовності *ph* та музичних ознак *m*, які задають нотну структуру та часову організацію співу. Загальна структура енкодера наведена на рис. 2.4.

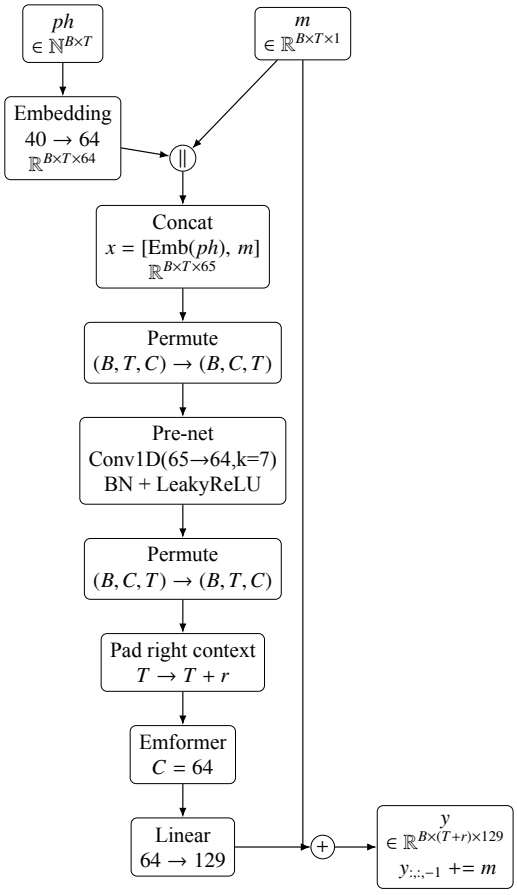

Рисунок 2.4 – Схема енкодера вхідних ознак (фонеми *ph* та музичні ознаки *m*).

Фонеми подаються у вигляді дискретних індексів і перетворюються на векторні представлення за допомогою шару вкладень. Музичні ознаки *m*, що

описують нотну структуру у часі, обробляються паралельно та об’єднуються з фонемними вкладеннями шляхом конкатенації по канальному виміру.

Отримане представлення x проходить через пренет на основі одномірної згортки з нормалізацією та нелінійною активацією, який виконує локальну часову агрегацію. Довші часові залежності моделюються за допомогою блока Emformer з урахуванням лівого та правого контексту.

На завершальному етапі лінійний шар проєктує ознаки у простір фіксованої розмірності, а резидуальне додавання компоненти m до останнього каналу забезпечує стабільне узгодження з нотними умовами. Вихід енкодера використовується як умовлювальний сигнал для наступних модулів моделі.

2.3.2 Модель часового зсуву

У синтезі співочого голосу часове узгодження між музичними нотами та фонемами відіграє критичну роль у збереженні природності та розбірливості згенерованого виконання. Дотримуючись підходу, запропонованого в [49], часовий зсув та тривалість ноти моделюються як дві окремі задачі прогнозування.

Часовий зсув визначає зміщення між моментом початку музичної ноти в нотному записі та фактичним початком відповідної фонеми у співочому виконанні. Таке зміщення виникає переважно через те, що приголосні звуки часто передують початку ноти, тоді як голосні зазвичай більш точно узгоджуються з її межами. Нехай g_n позначає еталонне значення часового зсуву для n-тої ноти, а $\hat{g}_n$ – відповідне прогнозоване значення. Модель навчається шляхом мінімізації середньоквадратичної похибки між прогнозованими та еталонними значеннями:

$$\mathcal{L}_{\text{lag}} = \mathbb{E}\|g - \hat{g}\|_2^2. \tag{2.25}$$

Під час синтезу прогнозований часовий зсув використовується для корекції моменту початку кожної ноти, формуючи скориговану ефективну тривалість $\hat{L}_n$, яка надалі застосовується для розподілу фонем.

2.3.3 Модель тривалості

Модель тривалості прогнозує довжину кожної фонеми в межах ноти на основі музичних та фонетичних ознак. Для n-тої ноти, що містить K_n фонем, модель генерує очікувані тривалості фонем $\hat{d}_{nk}$ за умови, що їх сумарна тривалість дорівнює скоригованій довжині ноти $\hat{L}_n$:

$$\sum_{k=1}^{K_n} \hat{d}_{nk} = \hat{L}_n. \tag{2.26}$$

Навчання мережі тривалостей здійснюється за критерієм середньоквадратичної похибки між прогнозованими та еталонними значеннями тривалостей фонем:

$$\mathcal{L}_{\text{dur}} = \mathbb{E}\|d - \hat{d}\|_2^2. \tag{2.27}$$

Зазначена цільова функція сприяє точному відтворенню фонемного таймінгу та забезпечує узгодженість прогнозованих тривалостей із музичним записом.

2.4 Висновки

У теоретичній частині роботи сформульовано та обґрунтовано метод узгодження потоку векторного поля для генеративного моделювання в латентному просторі. Розглянуто базову постановку flow matching, її модифіковані функції втрат та умовні варіанти, що дозволяють інтегрувати додаткові ознаки.

Запропоновано архітектуру генеративної моделі, яка поєднує умовне узгодження потоку з аудіо-автокодером, побудованим на основі DDSP-WORLD вокодера. Використання залишкової векторної квантизації дозволяє ефективно дискретизувати латентний простір та зменшити розмірність задачі генерації.

Також розроблено модель ознак музичної партитури, що включає енкодер

фонемних і нотних ознак, модель часового зсуву та модель тривалості. Це забезпечує коректне узгодження між лінгвістичною, музичною та акустичною інформацією у процесі синтезу співочого голосу.

РОЗДІЛ 3 ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

3.1 Набір даних

Для навчання моделі використовувався датасет Tohoku Kiritan [50], який містить 50 пісень загальною тривалістю приблизно 3,5 години. Перші три пісні були використані для тестування, наступні три – для валідації, а решта 44 – для навчання моделі.

Усі аудіосигнали були знижені до частоти дискретизації 24 кГц та нормалізовані до рівня −26 дБ. Датасет було сегментовано на фрагменти тривалістю 4 с, при цьому межі сегментації узгоджувалися з часовими мітками на рівні слів та природними паузами, що дозволило зберегти лінгвістичну та просодичну цілісність сигналу.

3.2 Оптимізація параметрів моделі

Навчання моделі здійснювалося з використанням оптимізатора Adam [51] з коефіцієнтом навчання 10^{-5}, параметрами $\beta_1 = 0.8$ та $\beta_2 = 0.999$ протягом 250 тис. кроків. Усі експерименти проводилися на одному графічному процесорі NVIDIA RTX 4060 з розміром пакета 16 та використанням обчислень у форматі з плаваючою комою половинної точності (fp16).

Навчання системи здійснювалося у декілька етапів. На початковому етапі компоненти моделі оптимізувалися окремо протягом перших 200 тис. кроків. Зокрема, автокодер навчався із використанням лише реконструктивних та квантизаційних функцій втрат. Адверсаріальний компонент для автокодера підключався після перших 20 тис. кроків його навчання, що дозволило стабілізувати спектральну реконструкцію та сформувати базову акустичну структуру сигналу перед застосуванням дискримінативного зворотного зв’язку.

Після завершення окремого навчання компонентів протягом наступних 50 тис. кроків здійснювалася спільна оптимізація всієї системи, включаючи енкодер, декодер, часові моделі, потоковий латентний предиктор та

дискримінатори. На цьому етапі модель навчалася в режимі End-to-End з використанням повної сукупності часових, реконструктивних, потокових та адверсаріальних функцій втрат, що забезпечило узгоджене тонке налаштування всіх модулів і підвищення перцептивної якості синтезованого співочого сигналу.

Загальна функція втрат визначається таким чином:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{lag}}\mathcal{L}_{\text{lag}} + \lambda_{\text{dur}}\mathcal{L}_{\text{dur}} + \lambda_{\text{STFT}}\mathcal{L}_{\text{STFT}} + \lambda_{\text{RVQ}}\mathcal{L}_{\text{RVQ}} + \lambda_{\text{GAN}}\mathcal{L}_{G} + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}} + \lambda_{\text{feature}}\mathcal{L}_{\text{feature}}, \tag{3.1}$$

де кожен коефіцієнт λ визначає вагу відповідної складової у загальній цільовій функції.

Вагові коефіцієнти функцій втрат було підібрано емпірично та встановлено як $\lambda_{\text{lag}} = 0.02$, $\lambda_{\text{dur}} = 0.02$, $\lambda_{\text{STFT}} = 1.0$, $\lambda_{\text{RVQ}} = 0.5$, $\lambda_{\text{GAN}} = 0.5$, $\lambda_{\text{FM}} = 1.0$ та $\lambda_{\text{feature}} = 0.1$. Зазначені значення обиралися на основі результатів валідації за сукупністю об’єктивних метрик: MCD та F_0-RMSE.

3.3 Методика оцінювання результатів

Оцінювання ефективності запропонованої моделі синтезу співочого голосу проводилось на відкладеній тестовій підмножині набору даних із використанням поєднання суб’єктивних та об’єктивних методів аналізу. Такий підхід дозволяє всебічно охарактеризувати як перцептивну якість синтезованого аудіо, так і точність відтворення спектральних і просодичних характеристик. Для порівняння використовувались базові системи ViSinger2 та HiddenSinger, а також еталонні записи та реконструкції автокодера.

3.3.1 Суб’єктивне оцінювання

Суб’єктивна оцінка якості проводилась за допомогою метрики Mean Opinion Score (MOS). Оцінювання виконувалось на підмножині з 80 синтезованих аудіозразків за участю п’яти незалежних слухачів. Кожному

учаснику пропонувалось оцінити перцептивну якість синтезованого співу за п’ятибальною шкалою, де вищі значення відповідають більш природному та приємному звучанню. Для підвищення надійності результатів MOS наводиться разом із 95% довірчими інтервалами.

3.3.2 Об’єктивне оцінювання

Об’єктивне оцінювання застосовувалось для аналізу спектральної та просодичної точності синтезу. Мел-кепстральне спотворення (Mel-Cepstral Distortion, MCD, дБ) використовується для кількісної оцінки відхилення спектральної форми між еталонним та згенерованим сигналами. Для аналізу висотної структури використовується середньоквадратична похибка основної частоти (F_0-RMSE), а точність визначення озвучених та неозвучених ділянок характеризується метрикою $V/UV\ F_1$. Як еталонні F_0 та V/UV характеристики використовуються значення, екстраговані з цільових аудіозаписів за допомогою вокодера WORLD [18].

3.3.3 Оцінювання обчислювальної ефективності

Для аналізу обчислювальної ефективності моделей додатково вимірювались показники реального часу синтезу (Real-Time Factor, RTF) та пікового використання пам’яті графічного процесора під час генерації. RTF визначається як відношення часу синтезу до тривалості згенерованого аудіо, а пікове використання пам’яті (у ГБ) характеризує максимальне навантаження на GPU під час інференсу. Ці показники дозволяють оцінити придатність запропонованого підходу до практичного та потенційно реального часу використання.

3.4 Аналіз експериментальних результатів

У таблиці 3.1 наведено перцептивні та акустичні метрики, які характеризують якість синтезу співочого голосу для всіх порівнюваних

систем. Середню суб’єктивну оцінку якості (Mean Opinion Score, MOS) подано з 95% довірчими інтервалами, тоді як інші метрики наведено у вигляді точкових оцінок.

Таблиця 3.1 – Показники якості синтезу на тестовому наборі

Модель	MOS	MCD (дБ)	F_0-RMSE	V/UV F_1
Еталонні дані	4.51 ± 0.07	0.00	0.00	1.000
Автокодер (реконструкція)	4.25 ± 0.10	1.47	15.03	0.968
ViSinger2	3.81 ± 0.08	2.71	19.74	0.951
HiddenSinger	3.49 ± 0.11	2.81	23.14	0.948
Запропонований метод (1 крок)	3.56 ± 0.07	2.92	23.52	0.937
Запропонований метод (8 кроків)	3.76 ± 0.13	2.78	22.89	0.949

Таблиця 3.2 містить показники обчислювальної ефективності розглянутих моделей, що характеризують ресурсні витрати та швидкодію під час синтезу.

Таблиця 3.2 – Обчислювальні характеристики моделей

Модель	К-сть параметрів	Пікове використання пам’яті (ГБ)	RTF
Автокодер (реконструкція)	3.2M	0.29	0.010 ± 0.002
ViSinger2	25.7M	4.07	0.021 ± 0.004
HiddenSinger	27.2M	3.82	0.784 ± 0.017
Запропонований метод (1 крок)	18.6M	1.94	0.016 ± 0.003
Запропонований метод (8 кроків)	18.6M	1.94	0.063 ± 0.003

3.5 Абляційне оцінювання

Для оцінювання внеску основних архітектурних компонентів було проведено два абляційні порівняння, узгоджені з поширеними альтернативами

у нейронному синтезі співочого голосу. У першому порівнянні зіставляється запропонований генератор на основі узгодження потоку з 8 кроками та генератор на основі дифузійної моделі, навчений за ідентичних умов щодо набору даних і механізмів умовлення. У другому порівнянні акустичний декодер на основі DDSP замінюється на поширені GAN-вокодери, зокрема SiFiGAN [22] та HiFiGAN [20].

Усі абляційні моделі навчалися з використанням одного й того самого набору даних, ідентичного розкладу оптимізації та однакових текстово-висотних умовних ознак, що забезпечує коректність порівняння. Оцінювання виконувалося на тій самій відкладеній тестовій підмножині, що й у основних експериментах.

У таблиці 3.3 наведено результати абляційного дослідження латентного предиктора.

Таблиця 3.3 – Абляція латентного предиктора (MOS)

Модель	**MOS**
Flow Matching (8 кроків)	3.76 ± 0.13
Diffusion (50 кроків)	3.32 ± 0.16

У таблиці 3.4 подано результати абляційного дослідження акустичного декодера.

Таблиця 3.4 – Абляція акустичного декодера (MOS)

Модель	**MOS**
DDSP-декодер	3.76 ± 0.13
SiFiGAN	3.73 ± 0.15
HiFiGAN	3.57 ± 0.09

3.6 Аналіз результатів дослідження

Результати, наведені в таблицях 3.1 та 3.2, демонструють, що запропонована модель забезпечує сприятливий компроміс між якістю синтезу та обчислювальною ефективністю. Порівняно з ViSinger2 та HiddenSinger,

запропонований підхід потребує на 28% менше параметрів і менш ніж удвічі менше пікового використання пам’яті графічного процесора під час інференсу, водночас зберігаючи порівнювану перцептивну якість. Зокрема, конфігурація з 8 кроками досягає значення MOS 3.76, що наближається до показника ViSinger2 (3.81), однак працює майже вдвічі швидше, про що свідчить нижче значення коефіцієнта реального часу (RTF), наведеного в таблиці 3.2. Однокрокова версія моделі забезпечує генерацію в реальному часі (RTF < 0.02), що підтверджує ефективність формулювання на основі узгодження потоку, яке дозволяє виконувати синтез лише за кілька ітеративних оновлень векторного поля швидкостей v_t, на відміну від сотень стохастичних кроків денойзингу, характерних для дифузійних моделей.

Об’єктивні метрики, наведені в таблиці 3.1, також свідчать про стабільну спектральну та просодичну точність. Незначне зростання показників F_0-RMSE та V/UV F_1 порівняно з базовими моделями вказує на незначні відхилення у висоті тону та визначенні озвученості, що може бути зумовлено зменшеною розмірністю латентного простору. Водночас нижче значення MCD для конфігурації з 8 кроками свідчить про покращену спектральну узгодженість та гармонійний баланс, особливо за умов багатокрокового уточнення, під час якого прогнозоване латентне представлення поступово наближається до цільового многовиду.

Хоча як дифузійні, так і потокові генеративні моделі використовують стохастичне семплювання, надмірна стохастичність іноді може призводити до варіативності висоти тону або таймінгу між запусками, що дещо знижує узгодженість результатів. Запропонований підхід зменшує цей ефект завдяки навчанню більш гладких векторних полів швидкостей у рамках узгодження потоку, що потребує значно меншої кількості кроків, ніж дифузійний денойзинг, водночас зберігаючи стохастичну гнучкість для керування експресивністю. Слід також зазначити, що оригінальна модель HiddenSinger [6] навчалася на значно більшому наборі даних (приблизно 150 годин аудіозаписів), тоді як у даній роботі оцінювання проводиться

на менших датасетах з метою аналізу здатності до узагальнення в умовах обмежених ресурсів. Ця різниця в масштабі даних, імовірно, є однією з причин спостережуваної різниці в суб'єктивній якості.

Висока швидкодія та конкурентна якість запропонованої системи зумовлені трьома ключовими архітектурними чинниками. По-перше, декодер на основі DDSP працює безпосередньо з компактними спектральними параметрами — спектральною огинаючою S_c та аперіодичністю A_c — замість повнорозмірних спектрограм, що знижує обчислювальні витрати без втрати перцептивної деталізації. По-друге, моделювання в латентному просторі суттєво зменшує розмірність прогнозування, дозволяючи мережі узгодження потоку навчати більш гладку динаміку з меншою кількістю параметрів. По-третє, цільова функція узгодження потоку забезпечує ефективну ітеративну генерацію шляхом оцінювання неперервних векторних полів швидкостей, а не виконання багатокрокового стохастичного денойзингу. Хоча внесок кожного з цих чинників окремо не було детально проаналізовано, їх сукупний ефект дозволяє досягти майже сучасного рівня якості синтезу за умови швидшої генерації та суттєво зменшеного використання пам'яті, що підтверджується результатами з таблиці 3.2.

Узгодженість формант та гармонічних траєкторій свідчить про здатність запропонованої моделі точно відтворювати тонку частотну еволюцію співочого голосу, доповнюючи кількісні результати, наведені в таблиці 3.1.

Результати абляційного дослідження латентного предиктора, подані в таблиці 3.3, показують, що генератор на основі узгодження потоку досягає вищої перцептивної якості порівняно з дифузійною альтернативою за ідентичних умов навчання, водночас уникаючи затримок, пов'язаних із багатокроковим семплюванням. Аналогічно, результати абляції акустичного декодера, наведені в таблиці 3.4, свідчать про те, що автокодер на основі DDSP демонструє перцептивну якість, порівнянну з SiFiGAN, та перевершує HiFiGAN.

3.7 Демонстрація роботи моделі

3.7.1 Аналіз роботи F_0-енкодера

На рисунку 3.1 показано логіт-представлення прогнозованої частоти основного тону F_0 у шкалі центів у часовому вимірі. Яскраво виражена жовта траєкторія відображає основну мелодичну лінію, тоді як затемнені області відповідають неозвученим або малоймовірним значенням F_0.

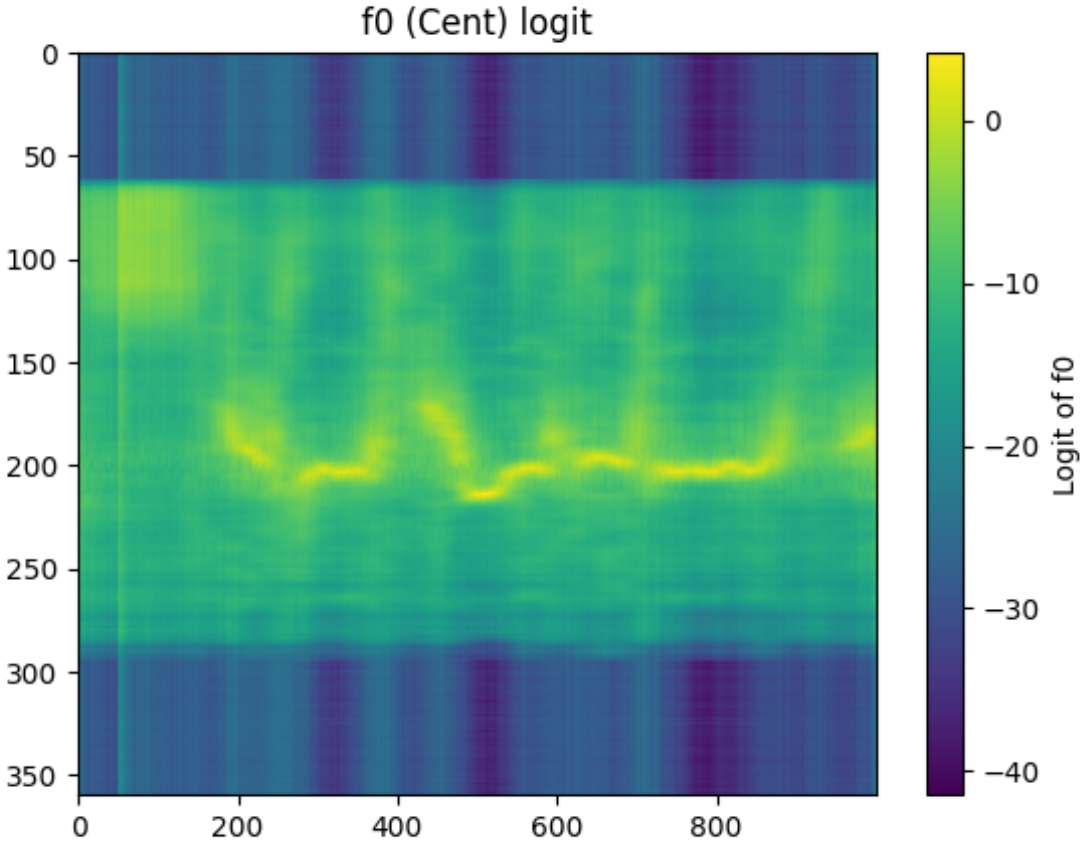

Рисунок 3.1 – Приклад логіт-представлення частоти основного тону

На рисунку 3.2 показано розподіл імовірностей для частоти основного тону F_0 у шкалі центів у часовому вимірі.

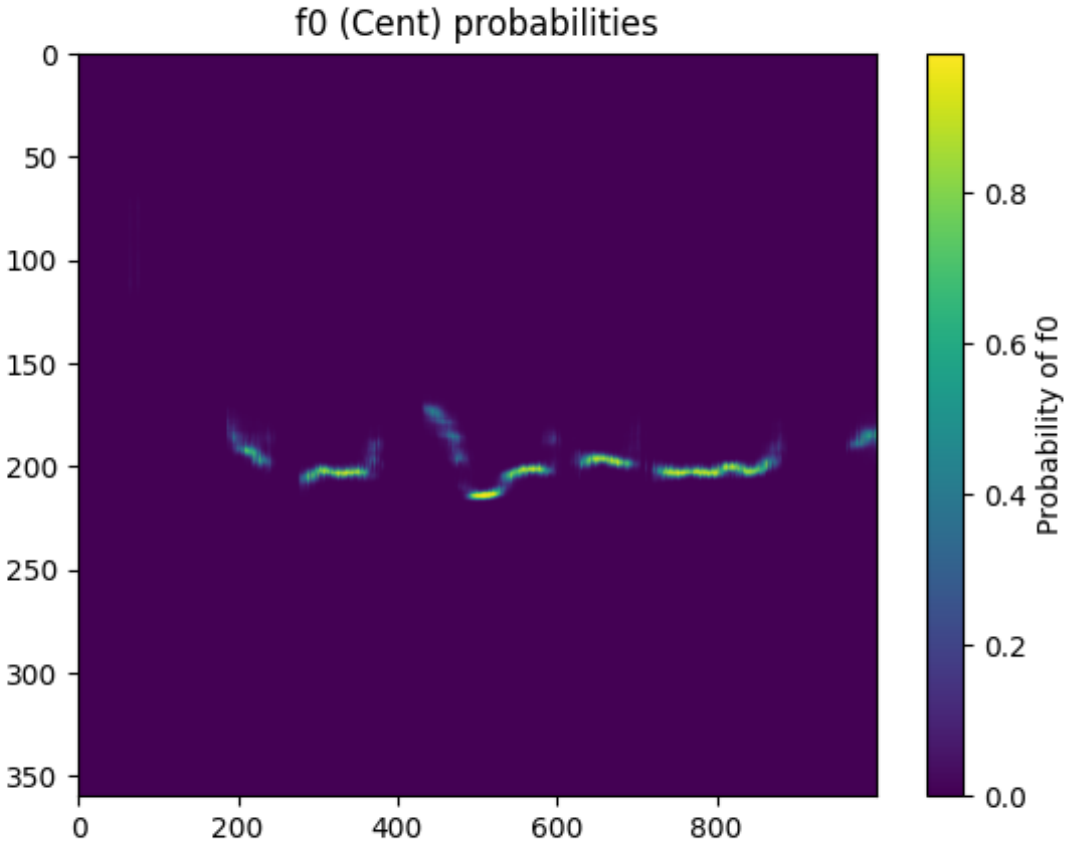

Рисунок 3.2 – Приклад передбачення ймовірності частоти основного тону

На відміну від логіт-представлення, дана візуалізація безпосередньо відображає концентрацію ймовірності навколо найбільш імовірних значень частоти основного тону. Вузька локалізована траєкторія свідчить про високу впевненість моделі в оцінці F_0 на озвучених ділянках, тоді як розсіяний або відсутній ймовірнісний масив відповідає неозвученим сегментам сигналу.

На рисунку 3.3 наведено мел-спектрограму еталонного аудіосигналу з накладеними траєкторіями частоти основного тону F_0 для еталонних та прогнозованих значень.

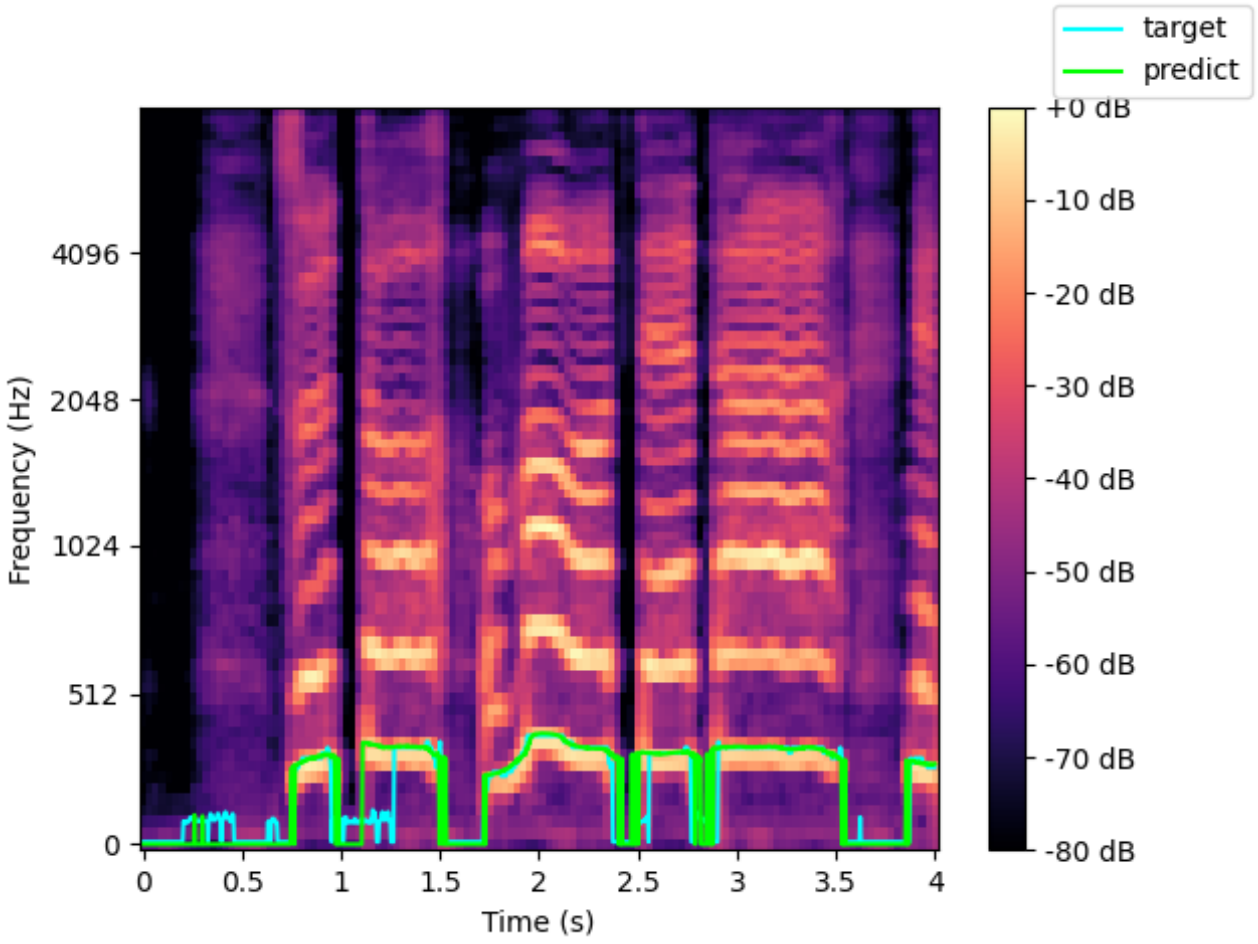

Рисунок 3.3 – Мел-спектрограма еталонного аудіосигналу з накладеними траєкторіями частоти основного тону F_0 для еталонних та прогнозованих значень.

Спектральна структура сигналу демонструє чітко виражені гармонічні компоненти та формантні області, тоді як накладені контури F_0 відображають часову динаміку висоти тону. Близька відповідність між еталонною та прогнозованою траєкторіями свідчить про коректне відтворення прогнозування в автоенкодері.

На рисунку 3.4 показано порівняння розподілів еталонних та прогнозованих значень частоти основного тону F_0, поданих у MIDI-шкалі. Криві густини відображають статистичну відповідність між цільовими та згенерованими значеннями F_0 у всьому діапазоні висот тону. Значна область перекриття між розподілами свідчить про близьку узгодженість прогнозованої частоти основного тону з еталонними даними.

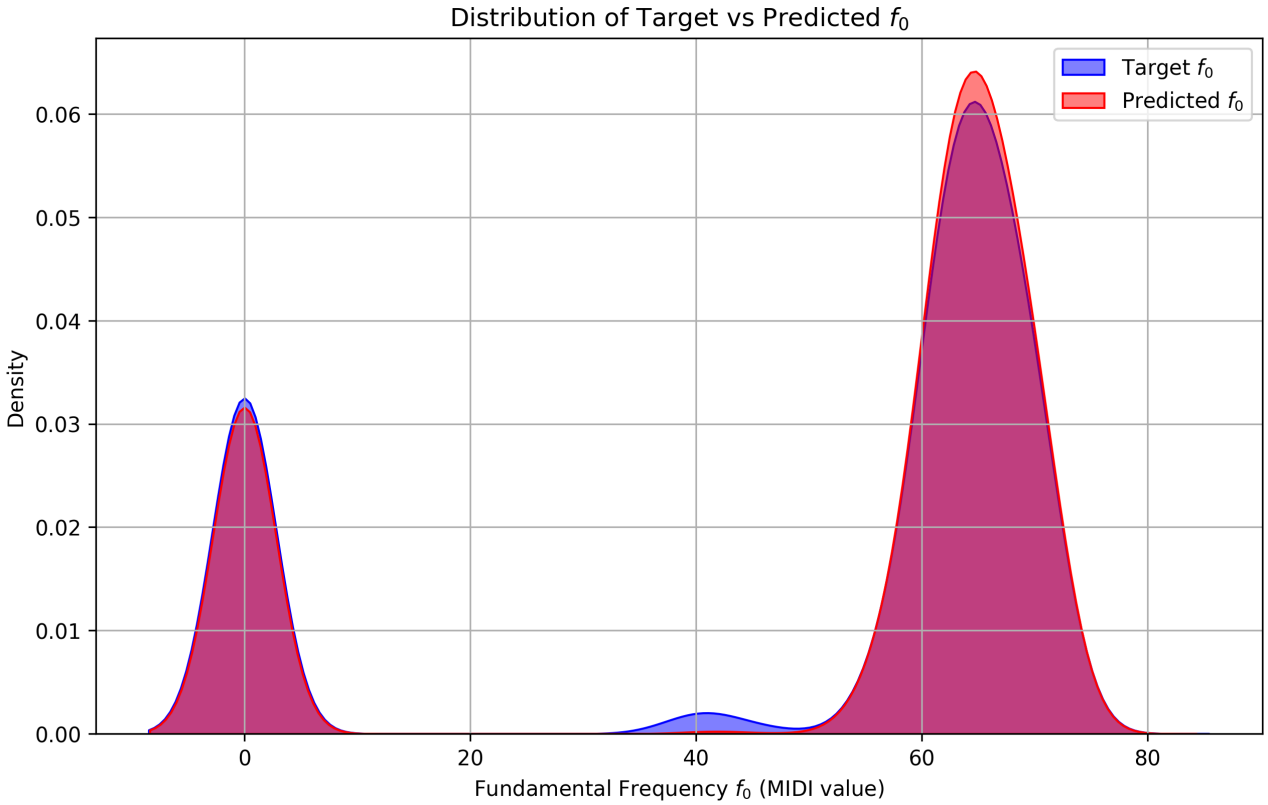

Рисунок 3.4 – Порівняння розподілів еталонних та прогнозованих значень частоти основного тону F_0 у MIDI-шкалі.

Аналіз розподілів показує, що модель коректно відтворює статистичні властивості частоти основного тону F_0 у межах усього діапазону висот тону, представленого в тестових даних. Основні модальні області прогнозованого розподілу добре узгоджуються з еталонними значеннями, що свідчить про відсутність систематичного зсуву у висоті тону.

Невеликі відмінності у формі розподілів, зокрема ослаблення або згладжування другорядних піків у прогнозованих значеннях, можуть бути пов’язані з сукупним впливом кількох чинників, включно з похибками моделі, обмеженнями алгоритму оцінювання частоти основного тону (WORLD) для еталонних даних, а також особливостями попередньої обробки та вирівнювання даних. Оскільки зазначені фактори є взаємопов’язаними, однозначне визначення джерела спостережуваних відмінностей виходить за межі даного аналізу.

Водночас висока ступінь перекриття розподілів підтверджує, що модель

зберігає коректний глобальний діапазон та статистичну структуру частоти основного тону.

На рисунку наведено матрицю невідповідностей (confusion matrix), що відображає відповідність між еталонними та прогнозованими значеннями частоти основного тону F_0, поданими у MIDI-шкалі в діапазоні від 52 до 77. Елементи на головній діагоналі відповідають випадкам точного відтворення значення MIDI, тоді як позадіагональні значення відображають помилки класифікації.

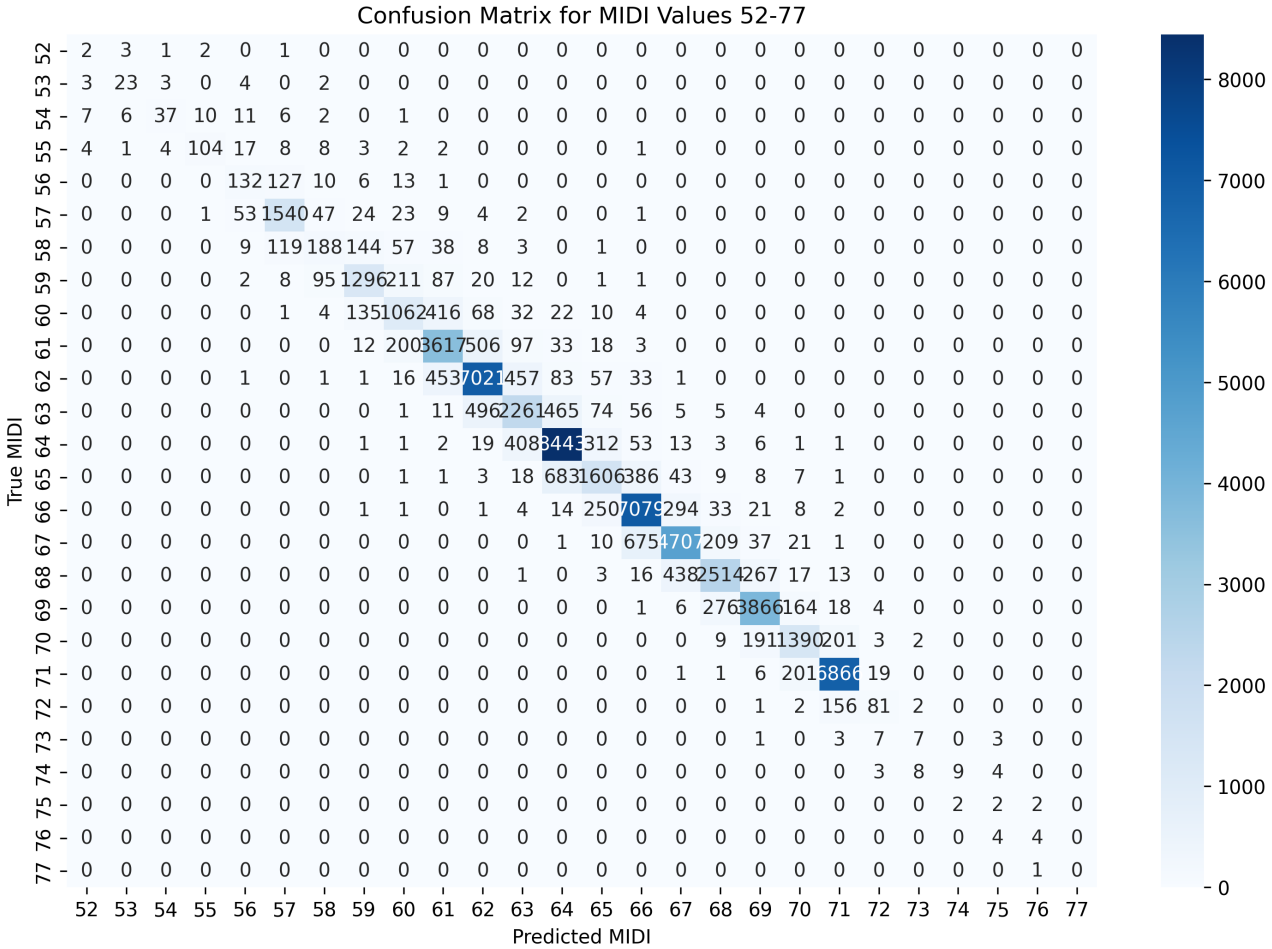

Рисунок 3.5 – Мел-спектрограма еталонного аудіосигналу з накладеними траєкторіями частоти основного тону F_0 для еталонних та прогнозованих значень.

Найбільша концентрація значень спостерігається вздовж головної діагоналі, що свідчить про переважно коректне прогнозування частоти основного тону. Помилки мають локальний характер і переважно обмежуються

сусідніми MIDI-значеннями, що відповідає незначним відхиленням у висоті тону.

3.7.2 Аналіз латентного простору та предиктора

На зображені 3.6 наведено цільове латентне представлення, отримане з енкодера аудіо-автокодера для вибраного фрагмента сигналу. Теплова карта відображає просторово-часову структуру латентних ознак, яка характеризується регулярними горизонтальними смугами та стабільною динамікою компонент у часі.

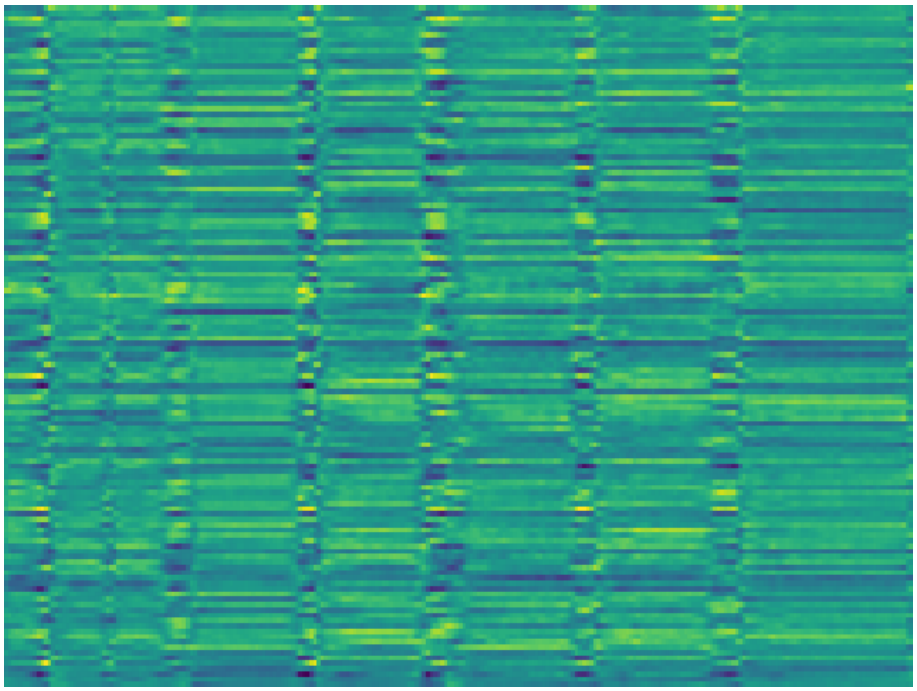

Рисунок 3.6 – Цільове латентне представлення *z*, отримане з енкодера автокодера

На зображенні 3.7 показано латентне представлення, передбачене моделлю узгодження потоку для того самого фрагмента. Передбачене подання загалом відтворює основну просторово-часову структуру цільового латентного простору, зберігаючи положення та протяжність домінантних компонент, хоча окремі локальні деталі є більш згладженими.

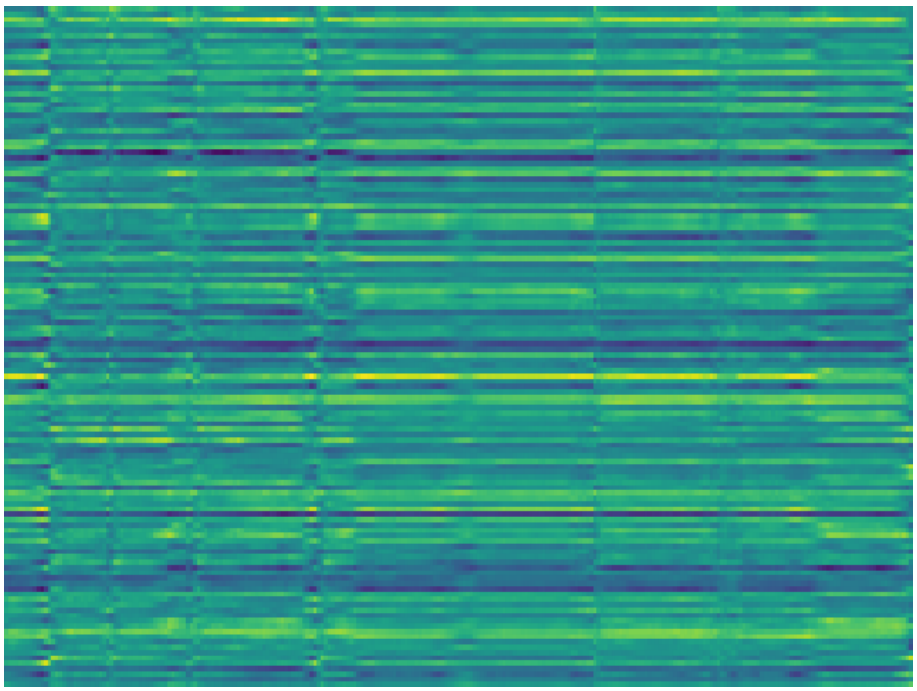

Рисунок 3.7 – Передбачене латентне представлення z_1, отримане за допомогою моделі узгодження потоку

Цільове представлення характеризується більш чіткою та стабільною горизонтальною структурою, що відображає узгоджені спектральні компоненти та регулярність латентної динаміки. Передбачене представлення загалом відтворює цю структуру: основні смуги зберігають положення та протяжність у часі, що свідчить про коректне наближення до цільового латентного розподілу.

Водночас у передбаченому варіанті спостерігається дещо підвищена локальна варіативність та згладжування окремих деталей, зокрема менш контрастні переходи між сусідніми смугами. Такі відмінності можуть бути наслідком апроксимаційних властивостей моделі узгодження потоку та обмеженої кількості ітерацій інтегрування.

На рис. 3.8 наведено еталонне латентне представлення z, що формується енкодером аудіо-автокодера. Воно характеризується відносно узгодженими часовими структурами (вертикальні зміни) та стабільними смугами енергії по «частотному»/канальному виміру, які відображають внутрішню організацію латентного простору для цього фрагмента.

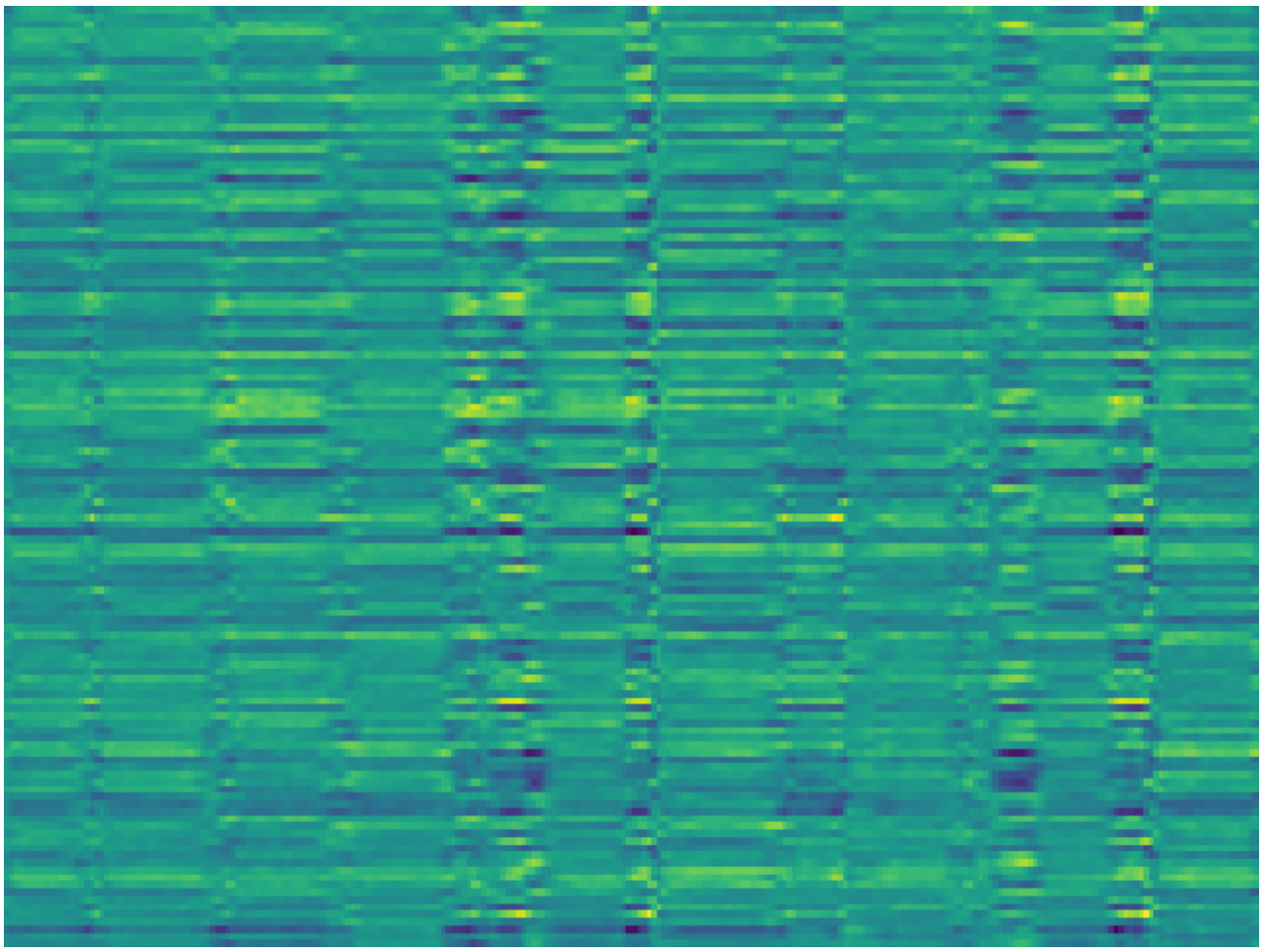

Рисунок 3.8 – Цільове латентне представлення, для прикладу з помилкою моделі узгодження потоку

На рис. 3.9 показано прогнозоване латентне представлення z_1, отримане предиктором на основі узгодження потоку. Візуально помітно суттєве відхилення від еталону: порушена структура смуг, з’являється нерегулярна «зернистість» та локальні контрастні сплески, а також слабша узгодженість часових патернів. Такий характер помилки свідчить не лише про невеликий зсув або згладжування, а про істотну невідповідність між передбаченим станом латенту та маніфолдом, на якому лежать коди, отримані енкодером.

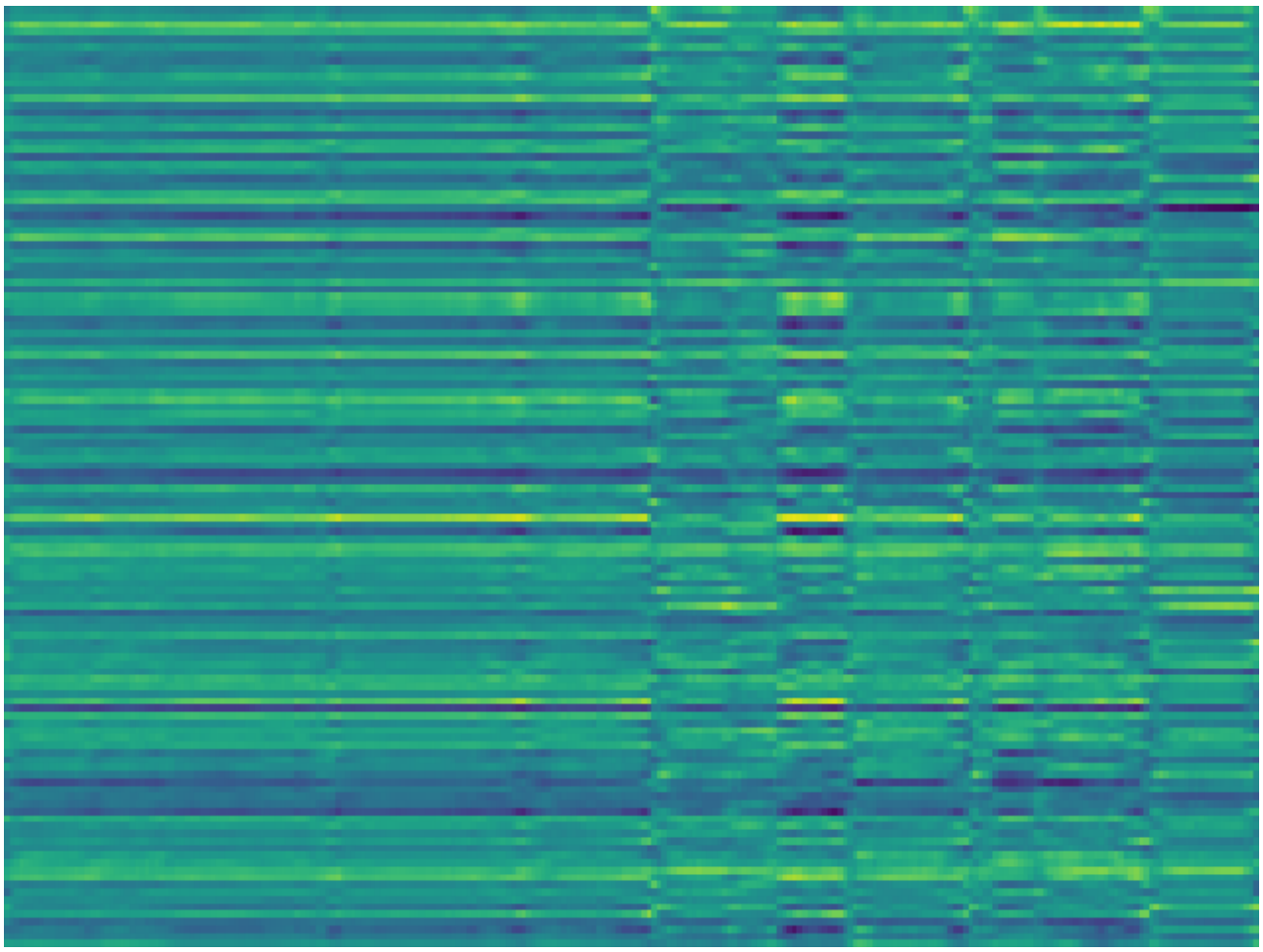

Рисунок 3.9 – Приклад латентного представлення, для якого модель узгодження потоку не змогла відтворити структуру цільового маніфолду

3.8 Висновки

У практичній частині роботи реалізовано запропонований метод синтезу співочого голосу та проведено експериментальне дослідження його ефективності. Описано процес підготовки набору даних, налаштування параметрів навчання та поетапну оптимізацію компонентів моделі.

За результатами об’єктивного та суб’єктивного оцінювання показано, що запропонований підхід забезпечує конкурентну якість синтезу порівняно з сучасними аналогами, при суттєво меншій обчислювальній складності. Абляційне дослідження підтвердило доцільність використання узгодження потоку в латентному просторі та ефективність DDSP-підходу для реконструкції аудіосигналу.

Візуалізація латентних представлень і аналіз роботи енкодера частоти

основного тону продемонстрували, що модель коректно відтворює основні закономірності цільового латентного простору, а виявлені відмінності мають локальний характер і не порушують загальної узгодженості синтезу.

РОЗДІЛ 4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1 Вплив шуму ультразвуку та інфразвуку на організм людини. Засоби захисту від шкідливої дії шуму.

4.1.1 Вплив шуму, ультразвуку та інфразвуку на організм людини

Шум, інфразвук та ультразвук належать до фізичних шкідливих виробничих факторів, дія яких за умов тривалої або інтенсивної експозиції може призводити до негативних змін функціонального стану організму людини. Відповідно до законодавства України у сфері охорони праці та санітарного благополуччя населення, рівні акустичних факторів на робочих місцях підлягають обов'язковому нормуванню та державному санітарному нагляду з метою запобігання професійним захворюванням і зниженню працездатності працівників [52, 53].

Шум є найбільш поширеним акустичним фактором виробничого середовища. Згідно з чинними санітарними нормами, він являє собою сукупність звукових коливань різної частоти та інтенсивності, що характеризуються рівнем звукового тиску, спектральним складом і тривалістю дії [54]. За часовими характеристиками шум поділяється на безперервний, переривчастий та імпульсний, а за спектральними – на широкосмуговий і тональний, що враховується під час його нормування та оцінки шкідливості.

Перевищення гранично допустимих рівнів шуму на робочому місці здатне призводити до специфічних уражень органу слуху – від тимчасового зсуву порогу чутності до розвитку професійної приглухуватості [54, 55]. Водночас навіть шум, рівень якого не перевищує встановлених нормативів, за умови тривалої дії може спричиняти неспецифічні функціональні порушення, зокрема підвищену втому, зниження концентрації уваги, дратівливість та порушення сну, що негативно впливає на ефективність розумової праці [56].

Фізіологічний механізм дії шуму не обмежується впливом на слуховий

аналізатор. Тривала акустична стимуляція призводить до перенапруження центральної нервової системи, активації вегетативних реакцій та змін серцево-судинної регуляції. За даними фахової літератури, такі зміни можуть формувати стан хронічної функціональної перевтоми навіть за відсутності виражених аудіологічних порушень [56].

У контексті розробки та тестування цифрових аудіотехнологій і систем синтезу вокалу джерелами акустичного навантаження є високопродуктивні робочі станції, серверне обладнання, а також процес тривалого моніторингу аудіосигналів через навушники. Особливістю такої діяльності є поєднання помірного рівня звукового тиску з тривалою експозицією та високими вимогами до точності слухового аналізу, що може призводити до слухової та когнітивної втоми [56].

Інфразвук – механічні коливання з частотою нижче 20 Гц – не сприймається слуховим аналізатором, проте чинить вплив на організм через вібраційні та резонансні механізми. Відповідно до санітарних норм, джерелами інфразвуку в виробничих та офісних приміщеннях можуть бути системи вентиляції й кондиціювання повітря, компресорне обладнання та інженерні комунікації будівель [54]. Тривала дія інфразвуку асоціюється з вегетативними порушеннями, головним болем, відчуттям дискомфорту та загальним зниженням працездатності.

Ультразвук – механічні коливання з частотою понад 20 кГц – у виробничих умовах зазвичай генерується спеціалізованим технологічним обладнанням. Згідно з ДСН 3.3.6.037-99, інтенсивна або тривала дія ультразвуку може спричиняти функціональні порушення з боку нервової та серцево-судинної систем, що обґрунтовує необхідність його нормування та контролю [54].

Окрім фізичних параметрів акустичного сигналу, важливе значення мають умови та режими його сприйняття. Всесвітня організація охорони здоров’я наголошує, що тривале прослуховування аудіосигналів через персональні аудіопристрої, зокрема навушники, за підвищеної гучності

або без достатніх перерв може створювати ризики для слуху та загального функціонального стану людини [57]. Це положення є актуальним для фахівців, діяльність яких пов’язана з тривалим аналізом аудіоматеріалу.

4.1.2 Засоби захисту від шкідливої дії акустичних факторів

Система захисту працівників від шкідливої дії шуму, інфразвуку та ультразвуку ґрунтується на ієрархічних принципах охорони праці, відповідно до яких пріоритет надається усуненню або зниженню рівня небезпечного фактора безпосередньо в джерелі його виникнення, а також застосуванню колективних засобів захисту перед індивідуальними [52]. Такий підхід спрямований на мінімізацію ризиків для здоров’я працівників незалежно від їх індивідуальних особливостей і рівня підготовки.

Інженерно-технічні заходи є основним способом зниження акустичного навантаження у виробничих, офісних та лабораторних приміщеннях. Вони спрямовані на зменшення рівня шуму та вібрацій у місцях їх виникнення або на шляху поширення. До таких заходів належать використання малошумного обчислювального й допоміжного обладнання, застосування звукопоглинальних кожухів і акустичних екранів, оздоблення приміщень звукоізоляційними та звукопоглинальними матеріалами, а також раціональне планування робочих зон. Зокрема, доцільною є локалізація серверного та інженерного обладнання в ізольованих приміщеннях або акустично відокремлених зонах [54, 58]. Для зменшення впливу інфразвуку та структурних вібрацій ефективним є застосування вібродемпфуючих основ і гнучких з’єднань у системах вентиляції та кондиціювання повітря, що відповідає вимогам санітарних норм.

Організаційні заходи відіграють особливо важливу роль для фахівців, діяльність яких пов’язана з аналізом аудіосигналів і не допускає повноцінного використання пасивних засобів індивідуального захисту органів слуху. Такі заходи спрямовані на обмеження експозиційного навантаження та раціональну організацію робочого процесу і включають:

1. регламентацію режимів праці та відпочинку з урахуванням

акустичного навантаження, зокрема впровадження періодичних перерв для розвантаження слухового аналізатора;

2. контроль рівня звукового тиску під час роботи з навушниками шляхом встановлення рекомендованих контрольних рівнів прослуховування;

3. запобігання впливу раптових акустичних піків за рахунок організаційних процедур безпеки та використання технічних засобів контролю сигналу у звуковому тракті.

Засоби індивідуального захисту органів слуху, зокрема протишумові вкладки та навушники, застосовуються у випадках, коли інженерно-технічні та організаційні заходи не забезпечують зниження рівня шуму до нормативних значень [59]. Водночас специфіка роботи з аудіоматеріалом полягає в необхідності точного слухового контролю якості сигналу, що істотно обмежує можливість використання класичних пасивних засобів захисту. У зв’язку з цим у таких умовах основний акцент переноситься на забезпечення акустичного комфорту робочого середовища та контроль тривалості й інтенсивності слухової експозиції.

Таким чином, забезпечення безпечних умов праці при роботі з аудіотехнологіями потребує комплексного підходу, який поєднує архітектурно-акустичні та інженерно-технічні рішення для зниження зовнішнього шуму, а також чітко регламентовані організаційні заходи для контролю внутрішнього акустичного навантаження, пов’язаного з професійною діяльністю.

4.2 Оцінка дії електромагнітного імпульсу (EMI) на елементи DDSP-WORLD вокодера і методи захисту

Безпека в надзвичайних ситуаціях є складовою цивільного захисту та спрямована на зменшення ризиків для персоналу, населення і технічних систем у разі аварій та техногенних впливів. Надзвичайні ситуації можуть супроводжуватися дією небезпечних фізичних факторів, зокрема пожеж, вибухів, токсичних викидів і електромагнітних впливів, що здатні порушувати

функціонування електронних та програмно-апаратних систем [60].

Відповідно до основ цивільної безпеки, підвищення стійкості об’єктів господарської діяльності в умовах надзвичайних ситуацій досягається шляхом поєднання організаційних, інженерно-технічних і програмних заходів, включаючи резервування, контроль критичних параметрів та забезпечення безпечного відновлення роботи після збоїв [61]. Застосування таких підходів є доцільним і для програмно-апаратних систем обробки сигналів, зокрема при оцінці їх стійкості до імпульсних збурень.

Сучасні системи синтезу мовлення та співочого мовлення реалізуються як програмно-апаратні комплекси та поєднують алгоритмічні методи DSP з глибокими нейронними мережами. Коректність їх функціонування визначається не лише точністю математичних моделей, але й стійкістю обчислювального процесу до зовнішніх фізичних впливів, зокрема електромагнітних завад імпульсного характеру.

4.2.1 Теоретичні засади впливу електромагнітного імпульсу та нормативно-правове забезпечення

Електромагнітний імпульс являє собою короткочасний нестаціонарний електромагнітний вплив, що характеризується різким зростанням напруженості поля та широким частотним спектром. Імпульсні електромагнітні завади здатні спричиняти раптові порушення у роботі електронних систем унаслідок індукції паразитних напруг і струмів у лініях живлення та передавання даних [62]. На рівні програмної реалізації такі впливи можуть проявлятися у вигляді короткочасних числових похибок, порушень синхронізації або збоїв у внутрішніх станах алгоритмів цифрової обробки сигналів, що призводить до деградації якості синтезу та появи чутних артефактів.

Стійкість електронних систем до дії електромагнітних імпульсів розглядається в межах концепції електромагнітної сумісності, яка визначає здатність обладнання функціонувати з установленою якістю в заданій електромагнітній обстановці без втрати працездатності [63]. В Україні вимоги

щодо електромагнітної сумісності обладнання встановлюються Технічним регламентом з електромагнітної сумісності обладнання, затвердженим постановою Кабінету Міністрів України № 1077 [63]. Практична реалізація цих вимог забезпечується застосуванням національних стандартів серії ДСТУ EN IEC 61000, які визначають загальні принципи та методи оцінки стійкості електронного обладнання до електромагнітних завад, у тому числі імпульсного характеру [64].

4.2.2 Формалізація завдання

У межах даного дослідження DDSP-WORLD вокодер розглядається як функціональний елемент програмно-апаратної системи обробки аудіосигналів, стабільність роботи якої може порушуватися під дією імпульсних збурень, еквівалентних за наслідками електромагнітним завадам.

Процес синтезу аудіосигналу описується відображенням

$$x(t) = \mathcal{V}(\mathbf{p}(t)), \tag{4.1}$$

де $\mathbf{p}(t)$ – вектор параметрів синтезу, що включає фундаментальну частоту, спектральну оболонку та параметри аперіодичності. Дія електромагнітного імпульсу інтерпретується як імпульсне збурення параметрів або внутрішніх станів алгоритму

$$\tilde{\mathbf{p}}(t) = \mathbf{p}(t) + \Delta\mathbf{p}(t). \tag{4.2}$$

Вихідний сигнал за наявності збурення визначається як

$$\tilde{x}(t) = \mathcal{V}(\tilde{\mathbf{p}}(t)), \tag{4.3}$$

а вплив імпульсного збурення оцінюється через відхилення

$$e(t) = \tilde{x}(t) - x(t). \tag{4.4}$$

Метою дослідження є оцінка впливу $\Delta\mathbf{p}(t)$ на стабільність і якість синтезу та визначення напрямів підвищення стійкості DDSP-WORLD вокодера.

4.2.3 Дослідження впливу імпульсних збурень на стабільність роботи DDSP-WORLD вокодера

Оцінка впливу електромагнітного імпульсу виконувалася як дослідження стійкості DDSP-WORLD вокодера до імпульсних збоїв, які за своїми наслідками є еквівалентними дії електромагнітних завад. Через відсутність доступу до спеціалізованої лабораторії електромагнітної сумісності дослідження не має характеру сертифікаційних ЕМС-випробувань і зосереджене на аналізі поведінки алгоритмічної частини та процесу виконання в умовах імпульсних порушень.

Дослідження проводилося у двох постановках. У першій постановці оцінювалася алгоритмічна чутливість до імпульсних збурень параметрів синтезу $\Delta\mathbf{p}(t)$, зокрема фундаментальної частоти, спектральної оболонки та фазових змінних. У другій постановці аналізувалася стійкість процесу виконання, коли імпульсні збої проявляються у вигляді числових помилок, порушень синхронізації.

Для кожного сценарію формувався еталонний сигнал $x(t)$ та сигнал $\tilde{x}(t)$, після чого аналізувалося відхилення $e(t)$, а також фіксувалися ознаки деградації якості синтезу, зокрема імпульсні артефакти та порушення періодичності. Додатково реєструвалися функціональні відмови системи.

Отримані результати показали, що найбільш чутливими до імпульсних збурень є параметри, пов’язані з формуванням періодичної складової сигналу, оскільки похибки в цих компонентах мають властивість накопичуватися. Це підтверджує доцільність застосування програмних механізмів обмеження та компенсації імпульсних похибок.

4.3 Висновки

У розділі з охорони праці та безпеки в надзвичайних ситуаціях розглянуто вплив акустичних факторів, зокрема шуму, ультразвуку та інфразвуку, на організм людини. Проаналізовано можливі ризики, пов’язані з використанням аудіосистем та засобів синтезу звуку.

Окремо досліджено вплив електромагнітного імпульсу на елементи DDSP-WORLD вокодера та розглянуто нормативно-правове забезпечення у сфері електромагнітної сумісності. Запропоновано рекомендації щодо підвищення стійкості програмно-апаратних компонентів до імпульсних збурень та забезпечення безпечних умов роботи.

ВИСНОВКИ

У кваліфікаційній роботі запропоновано модель для розв’язання науково-прикладну задачі синтезу співочого мовлення на основі глибинного навчання, спрямованого на досягнення високої якості синтезу за обмежених обчислювальних ресурсів. Основну увагу приділено поєднанню генеративного моделювання в латентному просторі з диференційовною цифровою обробкою сигналів у межах DDSP-WORLD вокодера.

У першій частині роботи виконано аналітичний огляд сучасних методів синтезу мовлення та синтезу співочого мовлення. Проаналізовано підходи на основі варіаційних, дифузійних та потокових моделей, а також сучасні TTS-моделі, що використовують узгодження потоку. Розглянуто існуючі вокодери та методи теселяції латентного простору, зокрема залишкове векторне квантування, що дозволило обґрунтувати вибір архітектурних рішень та актуальність дослідження.

У теоретичній частині роботи запропоновано метод узгодження потоку векторного поля для передбачення латентних представлень. Розглянуто модифіковані функції втрат, умовні варіанти узгодження потоку та механізми узгодження умовних ознак. Описано архітектуру генеративної моделі, аудіо-автокодера з залишковим векторним квантуванням та DDSP-WORLD декодером, а також архітектуру змагально-генеративної моделі. Запропоновано модель ознак музичної партитури, що включає енкодери фонемних і нотних ознак, модель часового зсуву та модель тривалості, що забезпечує коректне часово-музичне узгодження.

У практичній частині реалізовано запропонований метод та проведено експериментальні дослідження. Описано процес підготовки набору даних і оптимізації параметрів моделі. За результатами об’єктивного та суб’єктивного оцінювання показано, що запропонований підхід забезпечує конкурентну якість синтезу співочого голосу при значно меншій обчислювальній складності порівняно з дифузійними моделями. Абляційне дослідження

підтвердило ефективність використання узгодження потоку в латентному просторі та доцільність застосування DDSP-підходу для реконструкції аудіосигналу. Візуалізація результатів, зокрема аналіз роботи енкодера частоти основного тону, продемонструвала узгодженість передбачених характеристик з еталонними даними.

У розділі з охорони праці та безпеки в надзвичайних ситуаціях проаналізовано вплив шуму, ультразвуку та інфразвуку на організм людини, а також розглянуто вплив електромагнітного імпульсу на елементи DDSP-WORLD вокодера. Наведено нормативно-правове забезпечення та рекомендації щодо підвищення стійкості системи до шкідливих фізичних факторів і імпульсних збурень.

Отримані результати мають практичне значення та можуть бути використані для створення ефективних застосунків синтезу співочого голосу, зокрема в умовах обмежених обчислювальних ресурсів або малоресурсних наборів даних. Подальші дослідження можуть бути спрямовані на розширення підтримки мов і вокальних технік, покращення керованості синтезу та інтеграцію запропонованого методу у повноцінні музичні та мультимедійні системи.

СПИСОК ПОСИЛАНЬ

1. H. Kenmochi and H. Ohshita, “Vocaloid-commercial singing synthesizer based on sample concatenation.,” in *Interspeech*, vol. 2007, pp. 4009–4010, 2007.
2. K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, “Recent development of the hmm-based singing voice synthesis system-sinsy.,” in *SSW*, pp. 211–216, 2010.
3. P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, “Xiaoicesing: A high-quality and integrated singing voice synthesis system,” 2020.
4. Y. Ren, X. Tan, T. Qin, J. Luan, Z. Zhao, and T.-Y. Liu, “Deepsinger: Singing voice synthesis with data mined from the web,” 2020.
5. Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, “Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis,” 2022.
6. J.-S. Hwang, S.-H. Lee, and S.-W. Lee, “Hiddensinger: High-quality singing voice synthesis via neural audio codec and latent diffusion models,” *Neural Networks*, vol. 181, p. 106762, 2025.
7. J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, “Diffsinger: Singing voice synthesis via shallow diffusion mechanism,” 2022.
8. Y. Zhang, H. Xue, H. Li, L. Xie, T. Guo, R. Zhang, and C. Gong, “Visinger 2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer,” 2022.
9. Y. Zhang, W. Guo, C. Pan, D. Yao, Z. Zhu, Z. Jiang, Y. Wang, T. Jin, and Z. Zhao, “Tcsinger 2: Customizable multilingual zero-shot singing voice synthesis,” in *Findings of the Association for Computational Linguistics: ACL 2025*, p. 13280–13294, Association for Computational Linguistics, 2025.
10. W. Guo, Y. Zhang, C. Pan, R. Huang, L. Tang, R. Li, Z. Hong, Y. Wang, and Z. Zhao, “Techsinger: Technique controllable multilingual singing voice synthesis via flow matching,” 2025.

11. A. H. Liu, M. Le, A. Vyas, B. Shi, A. Tjandra, and W.-N. Hsu, “Generative pre-training for speech with flow matching,” 2024.
12. Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, “F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching,” 2025.
13. S. Mehta, R. Tu, J. Beskow, Éva Székely, and G. E. Henter, “Matcha-tts: A fast tts architecture with conditional flow matching,” 2024.
14. H. Lou, H. Paik, P. D. Haghighi, W. Hu, and L. Yao, “Latentspeech: Latent diffusion for text-to-speech generation,” 2024.
15. Y. Guo, C. Du, Z. Ma, X. Chen, and K. Yu, “Voiceflow: Efficient text-to-speech with rectified flow matching,” 2024.
16. H. Wang, S. Shan, Y. Guo, and Y. Wang, “Prosodyflow: High-fidelity text-to-speech through conditional flow matching and prosody modeling with large speech language models,” in *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7748–7753, 2025.
17. D. Griffin and J. Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.
18. M. Morise, F. Yokomori, and K. Ozawa, “World: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
19. H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, “Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds,” *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.
20. J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” 2020.
21. R. Yoneyama, Y.-C. Wu, and T. Toda, “Unified source-filter gan: Unified source-filter network based on factorization of quasi-periodic parallel wavegan,” *arXiv preprint arXiv:2104.04668*, 2021.

22. R. Yoneyama, Y.-C. Wu, and T. Toda, “Source-filter hifi-gan: Fast and pitch controllable high-fidelity neural vocoder,” 2023.
23. N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, “Efficient neural audio synthesis,” in *International Conference on Machine Learning*, pp. 2410–2419, PMLR, 2018.
24. A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.
25. A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, *et al.*, “Parallel wavenet: Fast high-fidelity speech synthesis,” in *International conference on machine learning*, pp. 3918–3926, PMLR, 2018.
26. R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617–3621, IEEE, 2019.
27. W. Ping, K. Peng, and J. Chen, “Clarinet: Parallel wave generation in end-to-end text-to-speech,” *arXiv preprint arXiv:1807.07281*, 2018.
28. J.-M. Valin and J. Skoglund, “Lpcnet: Improving neural speech synthesis through linear prediction,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5891–5895, IEEE, 2019.
29. J. Engel, L. Hantrakul, C. Gu, and A. Roberts, “Ddsp: Differentiable digital signal processing,” 2020.
30. S. Nercessian, “Differentiable world synthesizer-based neural vocoder with application to end-to-end audio style transfer,” 2023.
31. C.-Y. Yu and G. Fazekas, “Singing voice synthesis using differentiable lpc and glottal-flow-inspired wavetables,” *arXiv preprint arXiv:2306.17252*, 2023.
32. D.-Y. Wu, W.-Y. Hsiao, F.-R. Yang, O. Friedman, W. Jackson, S. Bruzenak, Y.-W. Liu, and Y.-H. Yang, “Ddsp-based singing vocoders: A new subtractive-based

synthesizer and a comprehensive evaluation," *arXiv preprint arXiv:2208.04756*, 2022.

33. P. Agrawal, T. Koehler, Z. Xiu, P. Serai, and Q. He, "Ultra-lightweight neural differential dsp vocoder for high quality speech synthesis," 2024.
34. B. Hayes, J. Shier, G. Fazekas, A. McPherson, and C. Saitis, "A review of differentiable digital signal processing for music and speech synthesis," *Frontiers in Signal Processing*, vol. 3, p. 1284100, 2024.
35. T. Bäckström, O. Räsänen, A. Zewoudie, P. P. Zarazaga, L. Koivusalo, S. Das, E. G. Mellado, M. Bouafif, and D. Ramos, "Introduction to speech processing," 2022.
36. A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.
37. A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Advances in neural information processing systems*, vol. 32, 2019.
38. P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.
39. N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.
40. Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," 2023.
41. S. Lee, Z. Lin, and G. Fanti, "Improving the training of rectified flows," *Advances in neural information processing systems*, vol. 37, pp. 63082–63109, 2024.
42. T. Luo, X. Miao, and W. Duan, "Wavefm: A high-fidelity and efficient vocoder based on flow matching," 2025.
43. B. Kim, Y.-G. Hsieh, M. Klein, M. Cuturi, J. C. Ye, B. Kawar, and J. Thornton, "Simple reflow: Improved techniques for fast flow models," *arXiv preprint arXiv:2410.07815*, 2024.

44. Y. Shi, Y. Wang, C. Wu, C.-F. Yeh, J. Chan, F. Zhang, D. Le, and M. Seltzer, “Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition,” 2020.
45. S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, “Bigvgan: A universal neural vocoder with large-scale training,” 2023.
46. J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “Crepe: A convolutional representation for pitch estimation,” 2018.
47. D.-Y. Wu, W.-Y. Hsiao, F.-R. Yang, O. Friedman, W. Jackson, S. Bruzenak, Y.-W. Liu, and Y.-H. Yang, “Ddsp-based singing vocoders: A new subtractive-based synthesizer and a comprehensive evaluation,” 2022.
48. X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” 2017.
49. Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Sinsy: A deep neural network-based singing voice synthesis system,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, p. 2803–2815, 2021.
50. I. Ogawa and M. Morise, “Tohoku kiritan singing database: A singing database for statistical parametric singing synthesis using japanese pop songs,” *Acoustical Science and Technology*, vol. 42, no. 3, pp. 140–145, 2021.
51. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.
52. Верховна Рада України, “Закон України «Про охорону праці».” Відомості Верховної Ради України, 1992. Дата звернення: 07.12.2025.
53. Верховна Рада України, “Закон України «Про забезпечення санітарного та епідемічного благополуччя населення».” Відомості Верховної Ради України, 1994. Дата звернення: 07.12.2025.
54. Міністерство охорони здоров’я України, “ДСН 3.3.6.037-99. Санітарні норми виробничого шуму, ультразвуку та інфразвуку,” 1999. Дата звернення: 07.12.2025.
55. Міністерство охорони здоров’я України, “Наказ № 540 від 23 березня 2023 р. Про затвердження граничних та робочих значень шумового впливу на робочому місці,” 2023. Дата звернення: 07.12.2025.

56. Грибан Г. В. and Негодченко О. В., *Охорона праці*. Центр учбової літератури, 2009.
57. World Health Organization, “Make listening safe,” 2021. Accessed: 07.12.2025.
58. Міністерство розвитку громад та територій України, “ДБН В.1.1-31:2013. Захист територій, будинків і споруд від шуму,” 2013. Дата звернення: 07.12.2025.
59. ДП «УкрНДНЦ», “ДСТУ en 458:2019. Засоби захисту органів слуху,” 2019. Дата звернення: 07.12.2025.
60. Стручок, В. С., *Безпека в надзвичайних ситуаціях: методичний посібник*. Тернопіль: ФОП Паляниця В. А., 2020. 156 с.
61. Стручок, В. С., *Техноекологія та цивільна безпека. Частина «Цивільна безпека»*. Тернопіль: ФОП Паляниця В. А., 2020. 156 с.
62. International Electrotechnical Commission, “Iec 61000 series: Electromagnetic compatibility (emc).”
63. Кабінет Міністрів України, “Технічний регламент з електромагнітної сумісності обладнання,” 2015. Постанова № 1077.
64. ДП «УкрНДНЦ», “ДСТУ en iec 61000-1-2:2022. Електромагнітна сумісність,” 2022.