

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»**

**Факультет прикладної математики
Кафедра системного програмування і
спеціалізованих комп'ютерних систем**

«На правах рукопису»
УДК 004.934

До захисту допущено:
Завідувач кафедри
_____ Віталій РОМАНКЕВИЧ
«__» _____ 2025 р.

**Магістерська дисертація
на здобуття ступеня магістра
за освітньо-професійною програмою «Системне програмування та
спеціалізовані комп'ютерні системи»
зі спеціальності 123 «Комп'ютерна інженерія»
на тему: «Спосіб розпізнавання нечіткого мовлення з використанням
нейромереж»**

Виконав:

студент II курсу, групи КВ-41мп

Шевченко Іван Іванович

Науковий керівник:

к.т.н, доцент

Потапова Катерина Романівна

Рецензент:

к.ф-м.н. доцент

Вовк Лілія Борисівна

Засвідчую, що у цій магістерській
дисертації немає запозичень з праць
інших авторів без відповідних
посилань.

Студент (-ка) _____

Київ – 2025 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет прикладної математики
Кафедра системного програмування і спеціалізованих комп'ютерних систем

Рівень вищої освіти – другий (магістерський)

Спеціальність – 123 «Комп'ютерна інженерія»»

Освітньо-професійна програма «Системне програмування та спеціалізовані комп'ютерні системи»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Віталій РОМАНКЕВИЧ

« ____ » _____ 2024 р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Шевченко Іван Іванович

1. Тема дисертації «Спосіб розпізнавання нечіткого мовлення з використанням нейромереж», науковий керівник дисертації Потапова Катерина Романівна, кандидат технічних наук, доцент, затверджені наказом по університету від «6» листопада 2025 р. №4836-с
2. Термін подання студентом дисертації 12 грудня 2025
3. Об'єкт дослідження: процес автоматичного розпізнавання нечіткого мовлення у системах штучного інтелекту
4. Вихідні дані: система розпізнавання нечіткого мовлення з використанням дифузійної моделі та варіаційного навчання.
5. Перелік завдань, які потрібно розробити:
 - проаналізувати сучасний стан проблеми нечіткого мовлення;
 - дослідити принципи роботи дифузійної моделі і можливості їх адаптації для реконструкції аудіосигналу;
 - реалізувати систему розпізнавання нечіткого мовлення, що поєднує дифузійну модель та варіаційне навчання;

- провести експериментальне оцінювання точності розпізнавання нечіткого мовлення та швидкодії розробленої системи.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу:

- презентація

7. Орієнтовний перелік публікацій:

– Тези доповіді «МОДИФІКОВАНИЙ ДИФУЗІЙНИЙ АЛГОРИТМ ПІДВИЩЕННЯ ТОЧНОСТІ СИСТЕМИ РОЗПІЗНАВАННЯ НЕЧІТКОГО МОВЛЕННЯ»;

– Тези доповіді «АНАЛІЗ ОБЧИСЛЮВАЛЬНОЇ СКЛАДНОСТІ ДИФУЗІЙНОГО АЛГОРИТМУ ПОПЕРЕДНЬОЇ ОБРОБКИ НЕЧІТКОГО МОВЛЕННЯ ДЛЯ СИСТЕМ РЕАЛЬНОГО ЧАСУ»;

– Стаття «МОДИФІКОВАНИЙ ДИФУЗІЙНИЙ АЛГОРИТМ ОБРОБКИ НЕЧІТКОГО МОВЛЕННЯ».

8. Дата видачі завдання 1 листопада 2024р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Вивчення літератури за темою магістерської дисертації	31.12.2024	
2	Робота над першим розділом магістерської дисертації; проведення наукового дослідження	31.03.2025	
3	Робота над другим розділом магістерської дисертації; розроблення програмного забезпечення;	30.06.2025	
4	Проведення наукового дослідження; робота над статтею за результатами дослідження	31.08.2025	
5	Робота над третім розділом магістерської дисертації	30.09.2025	
6	Завершення роботи над основною частиною магістерської дисертації	31.10.2025	

7	Оформлення текстової і графічної частини магістерської дисертації	1.12.2025	
---	--	-----------	--

Студент

Іван ШЕВЧЕНКО

Науковий керівник

Катерина ПОТАПОВА

РЕФЕРАТ

Актуальність теми. У сучасних системах розпізнавання мовлення особливою проблемою залишається розуміння нечітких або спотворених голосових сигналів, що характерно для реальних умов комунікації. Наявність фонового шуму, реверберації або нечіткої вимови знижує точність традиційних систем розпізнавання автоматичного мовлення. Дифузійна модель здатна відновлювати мовленнєвий сигнал шляхом поступового усунення шуму, що дозволяє підвищити точність розпізнавання нечіткого мовлення.

Мета роботи. Розробка системи розпізнавання нечіткого мовлення на основі модифікованого дифузійного алгоритму із використанням варіаційного навчання для підвищення точності розпізнавання мовлення.

Об'єкт дослідження. Процес автоматичного розпізнавання нечіткого мовлення у системах штучного інтелекту.

Предмет дослідження. Методи та алгоритми розпізнавання нечіткого мовлення.

Методи дослідження. У роботі застосовано: математичне моделювання процесу поширення шуму та його реконструкції; аналіз сучасних нейромережових архітектур (CNN, RNN, Transformer, Diffusion Models), порівняння; варіаційне навчання для персоналізації; експериментальне тестування на багатомовних аудіокорпусах із різними рівнями шуму; статистична оцінка точності та швидкодії системи.

Наукова новизна. Запропоновано модифікований дифузійний алгоритм попередньої обробки мовленнєвого сигналу. Обґрунтовано доцільність використання дифузійної моделі для підвищення точності розпізнавання нечіткого мовлення.

Практична цінність. Результати дослідження можуть бути використані для створення систем автоматичного розпізнавання мовлення, здатних працювати з нечітким мовленням у реальних умовах.

Апробація результатів. Основні положення та результати роботи були представлені на конференціях:

1. «Прикладна математика та комп'ютинг» ПМК-2025 та опубліковані у збірнику тез доповідей «МОДИФІКОВАНИЙ ДИFUЗІЙНИЙ АЛГОРИТМ ПІДВИЩЕННЯ ТОЧНОСТІ СИСТЕМИ РОЗПІЗНАВАННЯ НЕЧІТКОГО МОВЛЕННЯ»;
2. Науковий журнал «Таврійський науковий вісник. Серія: Технічні науки.» стаття «МОДИФІКОВАНИЙ ДИFUЗІЙНИЙ АЛГОРИТМ ОБРОБКИ НЕЧІТКОГО МОВЛЕННЯ» ;
3. V Міжнародна науково-практична конференція «OPEN SCIENCE NOWADAYS: MAIN MISSION, TRENDS AND INSTRUMENTS, PATH AND ITS DEVELOPMENT» тези «АНАЛІЗ ОБЧИСЛЮВАЛЬНОЇ СКЛАДНОСТІ ДИFUЗІЙНОГО АЛГОРИТМУ ПОПЕРЕДНЬОЇ ОБРОБКИ НЕЧІТКОГО МОВЛЕННЯ ДЛЯ СИСТЕМ РЕАЛЬНОГО ЧАСУ».

Структура та обсяг роботи. Магістерська дисертація складається з вступу, трьох розділів, висновків, список використаних джерел та додатків. Перший розділ присвячено теоретичним засадам розпізнавання спотвореного та нечіткого мовлення. Другий розділ містить опис модифікованого дифузійного алгоритму попередньої обробки аудіосигналу. Подано архітектуру запропонованої системи, описано процеси дифузійної реконструкції. Третій розділ містить практичну реалізацію системи розпізнавання нечіткого мовлення. Розглянуто вибір інструментів, формування датасету, процес навчання й тестування моделі. Подано інтерфейс користувача та результати тестування.

Ключові слова: нечітке мовлення, система автоматичного розпізнавання мовлення, дифузійна модель, CNN, RNN, Transformer, попередня обробка аудіосигналу.

ABSTRACT

Relevance of the topic. In modern speech recognition systems, one of the key challenges remains the understanding of unclear or distorted speech signals, which are typical of real communication conditions. The presence of background noise, reverberation, or unclear pronunciation reduces the accuracy of traditional automatic speech recognition systems. Diffusion models are capable of restoring speech signals through the gradual removal of noise, which makes it possible to improve the accuracy of recognizing unclear speech.

Purpose of the study. To develop an unclear speech recognition system based on a modified diffusion algorithm using variational learning to improve speech recognition accuracy.

Object of the study. The process of automatic recognition of unclear speech in artificial intelligence systems.

Subject of the study. Methods and algorithms for unclear speech recognition.

Research methods. The following methods were applied in this work: mathematical modeling of noise propagation and reconstruction processes; analysis and comparison of modern neural network architectures (CNN, RNN, Transformer, Diffusion Models); variational learning for personalization; experimental testing on multilingual audio corpora with different noise levels; statistical evaluation of system accuracy and computational performance.

Scientific novelty. A modified diffusion-based algorithm for speech signal preprocessing has been proposed. The feasibility of using a diffusion model to improve the accuracy of unclear speech recognition has been substantiated.

Practical significance. The research results can be used to develop automatic speech recognition systems capable of operating with unclear speech under real-world conditions.

Approbation of results. The main provisions and results of the study were presented at the following conferences:

1. *Applied Mathematics and Computing* (PMC-2025) and published in the conference proceedings as “A Modified Diffusion Algorithm for Improving the Accuracy of an Unclear Speech Recognition System”;
2. Scientific journal *Tavria Scientific Bulletin. Series: Technical Sciences*, article “A Modified Diffusion Algorithm for Unclear Speech Processing”;
3. V International Scientific and Practical Conference *OPEN SCIENCE NOWADAYS: Main Mission, Trends and Instruments, Path and Its Development*, proceedings “Analysis of the Computational Complexity of a Diffusion Algorithm for Unclear Speech Preprocessing in Real-Time Systems”.

Structure and scope of the thesis. The master’s thesis consists of an introduction, three chapters, conclusions, a list of references, and appendices. The first chapter is devoted to the theoretical foundations of distorted and unclear speech recognition. The second chapter describes the modified diffusion algorithm for audio signal preprocessing, presents the architecture of the proposed system, and explains the diffusion reconstruction processes. The third chapter focuses on the practical implementation of the unclear speech recognition system, including tool selection, dataset formation, model training and testing. A user interface and testing results are also presented.

Keywords: unclear speech, automatic speech recognition system, diffusion model, CNN, RNN, Transformer, audio signal preprocessing.

ЗМІСТ

СПИСОК ТЕРМІНІВ, СКОРОЧЕНЬ ТА ПОЗНАЧЕНЬ	10
ВСТУП	11
1. ТЕОРЕТИЧНІ ОСНОВИ РОЗПІЗНАВАННЯ НЕЧІТКОГО МОВЛЕННЯ. 13	
1.1. Поняття нечіткого мовлення та проблеми його автоматичного розпізнавання.....	13
1.2. Огляд сучасних систем розпізнавання мовлення (ASR).....	17
1.3. Використання штучних нейронних мереж у розпізнаванні аудіо	22
1.4. Теоретичні засади варіаційного навчання (Variational Learning, VAE) ...	26
Висновки до розділу 1	30
2. ДИФУЗІЙНА МОДЕЛЬ РОЗПІЗНАВАННЯ НЕЧІТКОГО МОВЛЕННЯ..	32
2.1. Принцип роботи дифузійної моделі та її адаптація для аудіо	32
2.2. Архітектура моделі	37
2.3. Використання варіаційного автоенкодера	42
2.4. Алгоритм розпізнавання нечіткого мовлення на основі запропонованої моделі.....	45
2.5. Оптимізація гіперпараметрів та процес навчання моделі.....	50
Висновки до розділу 2	55
3. РЕАЛІЗАЦІЯ СИСТЕМИ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВЛЕННЯ	57
3.1. Вибір інструментів і середовища реалізації (Python, PyTorch, Librosa тощо)	57
3.2. Формування словника та навчальних даних	62
3.3. Процес навчання акустичної моделі та тестування системи розпізнавання нечіткого мовлення за коротким словником	66
3.4. Аналіз результатів	74
Висновки до розділу 3	76
ВИСНОВКИ.....	77
СПИСОК ЛІТЕРАТУРИ.....	79

СПИСОК ТЕРМІНІВ, СКОРОЧЕНЬ ТА ПОЗНАЧЕНЬ

ASR (Automatic Speech Recognition) – автоматичне розпізнавання мовлення

GMM (Gaussian Mixture Model) – гаусівська сумішна модель

HMM (Hidden Markov Model) – прихована марковська модель

CNN (Convolutional Neural Network) – згорткова нейронна мережа

RNN (Recurrent Neural Network) – рекурентна нейронна мережа

U-Net – U-подібна архітектура згорткової нейронної мережі

CTC (Connectionist Temporal Classification) – алгоритм вирівнювання часових послідовностей без попереднього сегментування

LSTM (Long Short-Term Memory) – архітектура RNN з довгою пам'яттю

GRU – (Gated Recurrent Unit) – підвид рекурентної нейромережі

VAE (Variational AutoEncoder) – варіаційний автоенкодер

VAD (Voice Activity Detection) – детекція голосової активності

STT (Speech-to-Text) – перетворення мовлення у текст

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

WER (Word Error Rate) – частота помилок слів

CER (Character Error Rate) – частота помилок символів

SNR – (Signal-to-Noise Ratio) – відношення сигнал/шум

API – (Application Programming Interface) – спосіб взаємодії програм

ВСТУП

Мовлення є одним з основних засобів комунікації людини та відіграє важливу роль у взаємодії з інформаційними системами. У зв'язку з активним розвитком інтелектуальних технологій зростає потреба в автоматизації процесів обробки та інтерпретації мовленнєвих сигналів. Системи автоматичного розпізнавання мовлення (Automatic Speech Recognition, ASR) стали важливою складовою сучасних інформаційних рішень і широко застосовуються в голосових асистентах, системах автоматичного керування та інклюзивних технологіях.

Проте ефективність таких систем залежить від якості вхідного мовленнєвого сигналу та відповідності умов використання тим, на яких модель була навчена. У реальних умовах експлуатації мовлення часто є нечітким, що істотно знижує точність розпізнавання та обмежує практичну придатність ASR-систем [1].

Під нечітким мовленням розуміють мовленнєві сигнали, що містять різноманітні спотворення, зумовлені як акустичними, так і артикуляційними чинниками. До такої чинників належать фоновий шум, реверберація, низька якість запису, варіативність темпу та інтонації мовлення, а також індивідуальні особливості вимови. Окрему групу становить мовлення людей з порушеннями артикуляції, неврологічними захворюваннями, наслідками травм або операцій, для яких стандартні системи розпізнавання, орієнтовані на «еталонну» вимову, виявляються малоефективними або взагалі непридатними [1].

Актуальність проблеми нечіткого мовлення зростає у зв'язку з розширенням сфер застосування голосових інтерфейсів та зростанням потреби в інклюзивних технологіях. У сучасних умовах, зокрема в Україні, значна кількість користувачів стикається з порушеннями мовлення внаслідок неврологічних захворювань, травм або стресових розладів, що додатково

підкреслює важливість розробки ASR-системи, що здатна обробляти нечітке мовлення.

Традиційний підходи до автоматичного розпізнавання мовлення зазвичай базуються на класичних методах попередньої обробки сигналу та виділення ознак, зокрема мел-частотних кепстральних коефіцієнтів (MFCC) та енергетично нормованих кепстральних коефіцієнтів (PNCC). Хоча ці методи добре зарекомендували себе в контрольованих акустичних умовах, їх ефективність суттєво знижується за наявності значних спотворень сигналу [12].

Сучасні end-to-end моделі, зокрема трансформерні архітектури, демонструють високі показники точності, одна потребують надзвичайно великих обсягів навчальних даних і значних обчислювальних ресурсів, що обмежує їх використання в ресурсно-обмежених системах.

Отже, розробка та дослідження системи автоматичного розпізнавання нечіткого мовлення є актуальною задачею.

1. ТЕОРЕТИЧНІ ОСНОВИ РОЗПІЗНАВАННЯ НЕЧІТКОГО МОВЛЕННЯ

1.1. Поняття нечіткого мовлення та проблеми його автоматичного розпізнавання

Мовлення людини формується внаслідок узгодженої роботи дихальної, голосоутворювальної та артикуляційної систем організму. Потік повітря, що утворюється в легенях, проходить через голосову щілину, де внаслідок коливань голосових зв'язок виникає первинний звуковий сигнал, який характеризується певною основною частотою. Надалі цей сигнал зазнає змін у резонаторних порожнинах гортані, ротової та носової порожнин, а також під впливом положення та рухів артикуляційних органів, зокрема язика, губ і щелепи. Сукупність цих процесів визначає акустичні характеристики окремих мовних звуків (фонем) і їх взаємні переходи, що в результаті сприймається як безперервний мовленнєвий сигнал.

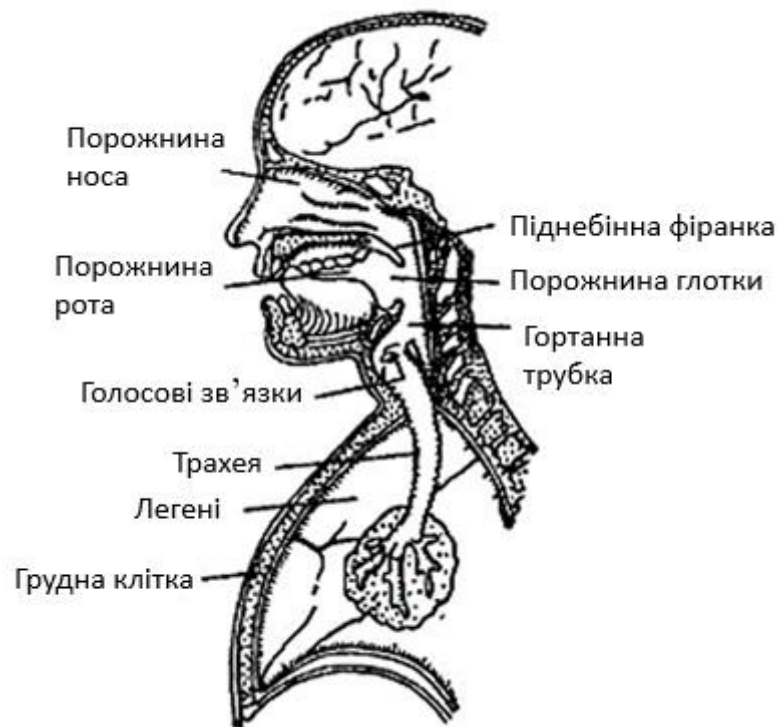


Рис. 1.1 – Анатомічна схема апарату мовотворення людини

Важливою характеристикою мовленнєвого сигналу є формантна структура, яка відображає резонансні властивості вокального тракту та

формується в процесі проходження голосового сигналу через артикуляційні порожнини. Форманти визначають спектральну структуру мовлення та є основою для розпізнавання голосних і приголосних звуків.

З фізичної точки зору мовлення можна розглядати як послідовність акустичних коливань тиску повітря, які поширюються в просторі у вигляді звукової хвилі. Такі коливання описуються набором параметрів, зокрема рівнем амплітуди, частотними характеристиками та спектральним складом, тобто розподілом енергії за частотами.

У реальних умовах мовленнєвий сигнал часто піддається впливу шуму, реверберації і нечіткої вимови. Для систем автоматичного розпізнавання мовлення це становить проблему, оскільки такі спотворення зменшують інформативність акустичних ознак, що є важливим для точного розпізнавання [22].

Нечіткість мовлення може мати акустичну природу, коли первинний мовленнєвий сигнал спотворюється умовами запису або поширення звуку. До таких спотворень належить шум, під яким розуміється будь-який сторонній акустичний сигнал, не пов'язаний безпосередньо з мовленням. Шум може бути як стаціонарним (наприклад, фоновий гул приміщення, робота вентиляційних систем), так і нестаціонарним (вуличні звуки, мовлення інших людей, технічні перешкоди). Наявність шуму призводить до зниження співвідношення сигнал/шум (SNR) та втрати інформативності мовленнєвого сигналу.

Окремим видом акустичних спотворень є реверберація, яка виникає тоді, коли мовлення багаторазово відбивається від різних поверхонь приміщення, і до мікрофона надходить не лише прямий сигнал, а й затримані відголоси.

Крім акустичних чинників, нечітке мовлення може бути зумовлене артикуляційною та просодичною варіативністю. До таких факторів належать швидкий або нерівномірний темп мовлення, нечітка артикуляція та індивідуальні особливості вимови. Оскільки більшість сучасних систем автоматичного розпізнавання мовлення орієнтовані на «еталонну» мову, за

наявності спотворень точність розпізнавання в таких системах істотно знижується [7].

Таблиця 1.1. Класифікація типів нечіткого мовлення та основні причини його виникнення

№	Тип нечіткості мовлення	Характеристика	Основні причини виникнення	Вплив на процес автоматичного розпізнавання
1	Акустична нечіткість	Зміна або спотворення звукової хвилі під впливом зовнішніх умов	Наявність шуму, реверберації, погана якість мікрофона, компресія сигналу, віддаленість мовця	Зниження співвідношення сигнал/шум, втрата фонетичних ознак, помилки у спектральному аналізі
2	Фонетична нечіткість	Неповна або неточна артикуляція звуків, редукція чи асиміляція фонем	Швидке мовлення, емоційне забарвлення, акценти, дефекти дикції, фізична втома мовця	Ускладнення сегментації, помилки у визначенні меж складів і слів
3	Просодична нечіткість	Варіації інтонації, темпу, гучності та наголосу	Стрес, емоційний стан, індивідуальні особливості мовлення, природні коливання темпу	Порушення часових залежностей у моделі, неправильне відтворення пауз і ритму
4	Лінгвістична нечіткість	Вживання неповних фраз, вставних слів, повторів, само переривань	Спонтанне мовлення, розмовний стиль, неформальні ситуації	Нестабільність послідовності слів, складність побудови граматичної структури
5	Індивідуальна нечіткість	Особистісні особливості мовця, що не збігаються з навчальними даними	Акцент, манера вимови, дефекти мовлення, логопедичні порушення	Зниження точності через відмінність від статистичних шаблонів моделі
6	Технічна нечіткість	Втрати або викривлення інформації під час оцифрування	Збої каналів зв'язку, спотворення при кодуванні, втрати пакетів у VoIP	Фрагментація аудіосигналу, помилки синхронізації,

		або передачі сигналу		часткове руйнування мовних ознак
--	--	-------------------------	--	-------------------------------------

Проблема нечіткого мовлення має широкий і зростаючий характер та охоплює різні вікові, соціальні й медичні групи населення. Згідно з науковими дослідженнями по всьому світу, кількість людей з порушеннями мовлення стабільно зростає протягом останніх років.

За даними міжнародних медичних та соціальних досліджень, від 10% до 12% дорослого населення мають ті чи інші порушення мовлення або комунікації. До цієї категорії належать особи з артикуляційними розладами, наслідками інсультів, черепно-мозкових травм, нейродегенеративних захворювань, а також люди з порушеннями слуху чи мовлення, спричиненими віковими змінами.

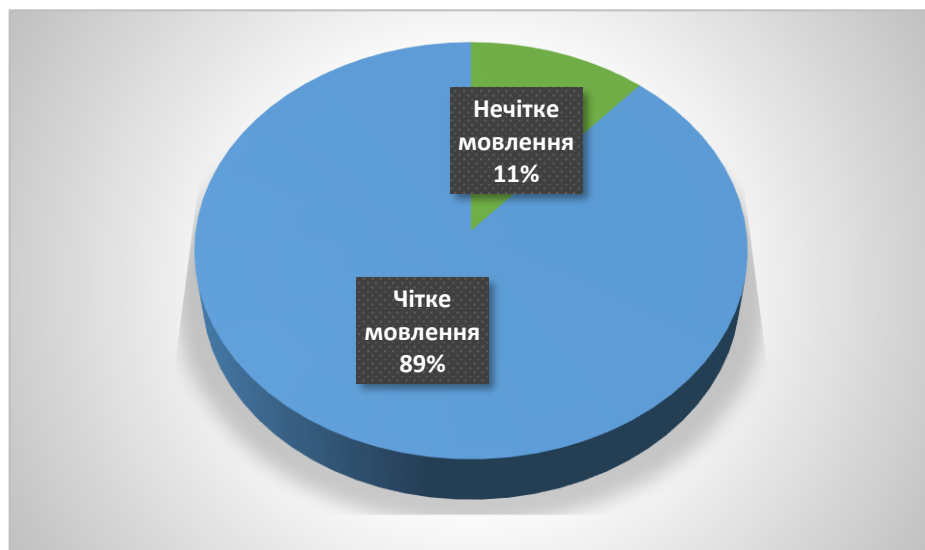


Рис 1.2 – частка дорослих людей з порушеннями мовлення серед дорослого населення у світі

Так за результатами National Health Interview Survey, частка дорослого населення США, що мають проблеми з комунікацією, зросла з 7,62% (17,89 млн дорослих у 2012 році) до 11,71% (29,92 млн у 2022 році).

Подібні тенденції спостерігаються і в інших країнах. Зокрема, у Великій Британії близько 7,6% дітей мають порушення мовлення, в Австралії цей

показник становить 3-6%, а у Швеції приблизно 16,9% дорослого населення мають ті чи інші комунікативні порушення. Такі дані підтверджуються, що проблема нечіткого мовлення є не локальною, а глобальною, і її масштаби мають стійку тенденцію до зростання.

Така тенденція зумовлена низкою факторів. Серед основних причин виокремлюють інсульты, неврологічні захворювання, черепно-мозкові травми, вікові зміни, а також наслідки стресових і травматичних подій, зокрема військових дій, які часто призводять до порушень артикуляції та втрати мовленнєвої чіткості.

Таким чином, з урахуванням наведених статистичних даних і тенденцій, задача автоматичного розпізнавання нечіткого мовлення є соціально значущою та науково актуальною.

1.2. Огляд сучасних систем розпізнавання мовлення (ASR)

Дослідження у сфері автоматичного розпізнавання мовлення мають тривалу історію та розвивалися паралельно з прогресом обчислювальної техніки, теорії сигналів і лінгвістики. У світовому масштабі перші системи розпізнавання з'явилися в середині XX століття та були здатні працювати лише з обмеженими словниками і «еталонним» мовленням одного диктора. Подальший розвиток методів динамічного вирівнювання часу, прихованих марковських моделей, а згодом і глибоких нейронних мереж суттєво розширив можливості ASR-систем, дозволивши перейти до розпізнавання злитого мовлення, великих словників і багатомовних сценаріїв.

В Україні наукові дослідження з автоматичного розпізнавання мовлення розпочалися ще у 1960-х роках і формувалися як окремий напрям у межах кібернетики та обробки сигналів. Значний внесок у становлення цієї галузі зроблено науковими колективами Інституту кібернетики АН УРСР, а згодом Міжнародного науково-навчального центру інформаційних технологій та систем НАН і МОН України [1]. Саме тут були закладені теоретичні основи

генеративного підходу до розпізнавання мовлення, відомого як ієрархічний композиційно-динамічний підхід, який ґрунтується на моделюванні процесів породження мовленнєвого сигналу та його порівнянні з еталонними структурами за допомогою динамічного програмування [1].

Одним із ключових результатів цих досліджень стало формування генеративного підходу до розпізнавання мовлення, відомого як ієрархічний композиційно-динамічний підхід. Даний підхід ґрунтується на моделюванні процесів породження мовленнєвого сигналу на різних рівнях – від акустичного до лінгвістичного – та подальшому зіставленні спостережуваного сигналу з еталонними структурами за допомогою методів динамічного програмування [1]. Така концепція дозволила формалізувати процес розпізнавання як задачу оптимального вирівнювання часових послідовностей і стала важливим теоретичним підґрунтям для подальшого розвитку як статистичних, так і нейромережевих методів розпізнавання мовлення [1].



Рис. 1.3 – Експериментальна система розпізнавання злитого мовлення на ЕОМ “БЭСМ-6” (1970)

Упродовж наступних десятиліть в Україні було створено низку експериментальних і прикладних систем розпізнавання мовлення, зокрема системи усного діалогу, голосового керування та багатомовні платформи,

орієнтовані на українську мову. Важливою особливістю цих розробок була увага до лінгвістичної специфіки української мови, включно з її фонемною структурою, інтонаційними особливостями та коартикуляційними ефектами, що суттєво впливають на акустичну реалізацію мовлення [1].

Сучасні системи автоматичного розпізнавання мовлення характеризуються широким застосуванням глибинних мереж та переходом від класичних архітектур до end-to-end підходів. Провідні технологічні компанії та наукові спільноти розробляють як комерційні хмарні рішення, так і відкриті платформи, орієнтовані на дослідницьке використання.

Хмарні сервіси, такі як Google Speech-to-Text, Microsoft Azure Speech Services та Amazon Transcribe, базуються на масштабованих нейронних архітектурах, зокрема Conformer, Transformer-Transducer та гібридних CTC/Attention моделях. Вони забезпечують високу точність розпізнавання, підтримку десятків і навіть сотень мов, а також можливість потокового розпізнавання в режимі реального часу. Завдяки потужній обчислювальній інфраструктурі такі системи демонструють низькі значення метрики Word Error Rate (WER), однак їх використання пов'язане з необхідністю передачі даних у хмару та залежністю від зовнішніх сервісів.

Окрему групу становлять універсальні end-to-end моделі, зокрема Whisper від OpenAI та Wav2Vec 2.0 від Meta. Whisper вирізняється високою стійкістю до шуму, багатомовністю та здатністю виконувати як розпізнавання, так і переклад мовлення, що робить її популярною у дослідницьких і прикладних задачах. Wav2Vec 2.0, у свою чергу, використовує підхід самонавчання без повних транскрипцій, що дозволяє досягати високої точності навіть за обмеженої кількості розмічених даних і є особливо актуальним для мов з обмеженими ресурсами.

Важливу роль у наукових дослідженнях відіграють відкриті платформи, такі як Kaldi, DeepSpeech 2 та SpeechBrain. Kaldi є модульним фреймворком, який підтримує як класичні GMM-HMM, так і нейромережеві DNN-HMM архітектури, забезпечуючи максимальну гнучкість налаштування, проте

вимагаючи значних зусиль для конфігурації. DeepSpeech 2 та SpeechBrain орієнтовані на спрощення процесу розробки та інтеграції нейромережових ASR-систем, що робить їх популярними в академічному середовищі.

Таблиця 1.2. Порівняльна характеристика сучасних систем розпізнавання мовлення (ASR)

№	Назва системи	Розробник / Платформа	Архітектура моделі	Ключові особливості	Точність розпізнавання (WER, %)	Сфера застосування
1	Google Speech-to-Text	Google Cloud	Енд-ту-енд Conformer + Transducer	Підтримка 120+ мов, автоматична сегментація, потужна хмарна інфраструктура, реальний час	4,5–6,0	Віртуальні асистенти, стенограми, корпоративні сервіси
2	Microsoft Azure Speech Services	Microsoft	Transformer-Transducer, Language Fusion	Інтеграція з Office, персоналізація словників, адаптація до мовця	5,0–6,5	Багатомовні системи, медичні й освітні додатки
3	Amazon Transcribe	Amazon Web Services	Hybrid CTC/Attention	Підтримка потокового розпізнавання, автоматичне видалення шумів, створення субтитрів	6,0–8,0	Бізнес-аналітика, контакт-центри, медіа
4	DeepSpeech 2	Baidu Research	Bidirectional RNN + CTC	Відкрите ПЗ, простота інтеграції, висока швидкість обробки	7,0–9,0	Академічні дослідження, офлайн-обробка мовлення
5	Whisper	OpenAI	Encoder-Decoder Transformer	Робастність до шуму, розпізнавання	4,7–6,3	Мультимовні системи, дослідження,

			(Multilingual)	90+ мов, здатність до перекладу, без хмарних обмежень		медіа-аналітика
6	Kaldi	Open Source Community	GMM-HMM, DNN-HMM (модульна архітектура)	Гнучкість конфігурацій, вимогливість до налаштування, можливість локального навчання	8,0–12,0	Наукові експерименти, невеликі корпоративні рішення
7	Wav2Vec 2.0	Meta (Facebook AI)	Self-Supervised Transformer	Навчання без транскрипцій, висока точність при мінімальних даних, адаптація до шуму	3,8–5,0	Нейромережі, дослідження, інтеграція у власні системи
8	SpeechBrain	Open Source / PyTorch	Модульна E2E Transformer система	Об'єднання моделей для ASR, TTS, speaker ID, open-source бібліотека	5,0–7,5	Академічні, промислові застосування
9	Apple Siri ASR	Apple	Proprietary Neural Transducer	Висока точність у шумних середовищах, локальне оброблення на пристрої	4,0–5,5	Мобільні пристрої, розумні асистенти

Загалом аналіз сучасних ASR-систем свідчить про те, що висока точність розпізнавання досягається за рахунок масштабності моделей, великих обсягів навчальних даних і потужних обчислювальних ресурсів. Водночас такі рішення часто є складними, дорогими або залежними від хмарної інфраструктури. Це підкреслює актуальність досліджень, спрямованих на підвищення стійкості систем автоматичного розпізнавання мовлення до

нечіткої вимови та акустичних спотворень із використанням більш компактних і ресурсно-ефективних підходів, зокрема шляхом попередньої обробки сигналу за допомогою дифузійних моделей.

1.3. Використання штучних нейронних мереж у розпізнаванні аудіо

Розвиток систем автоматичного розпізнавання аудіосигналів, зокрема мовлення, тісно пов'язаний із впровадженням методів штучного інтелекту та глибинного навчання. Штучні нейронні мережі забезпечили суттєвий прогрес у точності та стійкості аудіоаналізу, дозволивши автоматично моделювати складні нелінійні залежності між акустичними ознаками та мовними одиницями без необхідності ручного проектування правил.

Початкові підходи до використання нейронних мереж у розпізнаванні мовлення базувалися на класичних нейронних мережах прямого типу [1], які застосовувалися для класифікації фонем або акустичних станів на основі заздалегідь виділених ознак, таких як MFCC. Проте обмежена здатність таких моделей враховувати часову структуру сигналу стримувала їх ефективність. Подальший розвиток привів до застосування рекурентних нейронних мереж (RNN), зокрема архітектур LSTM та GRU, які дозволяють моделювати часові залежності та контекст мовлення на довгих інтервалах.

Значним кроком уперед стало впровадження згорткових нейронних мереж (CNN) для аналізу спектральних представлень аудіосигналів, таких як мел-спектрограми. CNN здатні автоматично виявляти локальні часово-частотні структури, що відповідають формантним структурам, переходам між фонемами та іншим інформативним акустичним ознакам. Поєднання CNN із рекурентними мережами дозволило створювати гібридні архітектури, які ефективно поєднують локальний спектральний аналіз із глобальним часовим контекстом.

Сучасні системи розпізнавання аудіо дедалі частіше базуються на трансформерних архітектурах, які використовують механізми самоуваги для

моделювання залежностей між різними фрагментами сигналу незалежно від їх відстані у часі. Це дозволяє більш точно враховувати контекст мовлення та зменшувати втрати інформації, характерні для рекурентних моделей. Трансформери стали основою багатьох сучасних end-to-end ASR-систем, включно з Conformer, Transformer-Transducer та багатомовними моделями.

Таблиця 1.3. Порівняння підходів до розпізнавання нечіткого мовлення з використанням нейронних мереж

№	Підхід / Архітектура	Основна ідея методу	Переваги використання	Недоліки / Обмеження	Сфери практичного застосування
1	Convolutional Neural Networks (CNN)	Аналіз спектрограм і виявлення локальних частотно-часових ознак у мовленні	Висока швидкість обробки, стійкість до шумів, добре розпізнає короткі звукові патерни	Недостатня здатність до моделювання довготривалих залежностей, чутливість до акцентів	Обробка коротких команд, розпізнавання ключових слів, голосовий пошук
2	Recurrent Neural Networks (RNN, LSTM, GRU)	Послідовне моделювання мовлення з урахуванням часових залежностей	Ефективне врахування контексту, плавна реконструкція фраз, добре працює з послідовними даними	Повільне навчання, проблеми з довгими послідовностями, деградація градієнтів	Транскрибування аудіо, автоматичне субтитрування, голосові асистенти
3	Transformer / Conformer	Використання механізму самоуваги для одночасного	Висока точність, паралельність обчислень, стабільність до	Велика кількість параметрів, значне споживання	Мультимовні системи розпізнавання, хмарні ASR-сервіси,

		аналізу всього аудіосигналу	шумів, масштабованість	ресурсів, складність налаштування	переклад аудіо в текст
4	Generative Diffusion Models	Покрокове очищення сигналу від шуму та реконструкція спотвореного мовлення	Здатність відновлювати нечіткі або частково втрачені дані, природність звучання	Велика обчислювальна складність, складність оптимізації дифузійного процесу	Розпізнавання нечіткого мовлення, покращення якості аудіо, відновлення голосових даних
5	Variational Autoencoder (VAE)	Створення латентного простору для моделювання індивідуальних особливостей голосу	Підтримка персоналізації, можливість адаптації до окремого користувача	Може втрачати дрібні фонетичні деталі, ризик перенавчання при малих вибірках	Персоналізовані голосові інтерфейси, адаптація до акцентів, допомога людям із мовленнєвими порушеннями
6	Hybrid Neural Systems (CNN + Transformer)	Поєднання локальної чутливості CNN і глобального контексту Transformer	Висока точність, баланс швидкості й стійкості, здатність працювати в реальному часі	Необхідність великої кількості даних і обчислювальних ресурсів	Реальні голосові асистенти, потокове розпізнавання мовлення, мультимодальні системи

Подані в таблиці 1.3 підходи до розпізнавання нечіткого мовлення відрізняються передусім рівнем, на якому здійснюється компенсація акустичних і артикуляційних спотворень. Частина методів орієнтована на аналіз локальних часово-частотних ознак мовлення, інші зосереджуються на моделюванні довготривалого контексту, тоді як генеративні підходи

виконують відновлення сигналу або адаптацію системи до індивідуальних особливостей мовця ще до етапу декодування.

Згорткові нейронні мережі ефективно працюють із спектральними представленнями мовлення, виявляючи стійкі локальні частотно-часові структури. Це дозволяє їм зберігати відносну стійкість до помірного шуму та забезпечувати високу швидкість обробки. Водночас обмежена здатність таких моделей враховувати довготривалі часові залежності знижує їх ефективність у випадках спонтанного мовлення, сильної реверберації або значної варіативності вимови.

Рекурентні нейронні мережі, зокрема архітектури LSTM і GRU, забезпечують більш повне врахування часової структури мовлення, що є важливим для відновлення цілісності фраз у присутності спотворень. Вони здатні використовувати контекст для часткової компенсації нечіткої артикуляції та шумових завад, однак їх практичне застосування ускладняється повільним навчанням і зниженням ефективності при роботі з дуже довгими послідовностями.

Трансформерні архітектури та їх модифікації, зокрема Conformer, демонструють найвищі показники точності завдяки використанню механізму самоуваги, який дозволяє одночасно аналізувати весь аудіосигнал. Це забезпечує високу стійкість до шуму та реверберації і робить такі моделі особливо придатними для розпізнавання нечіткого мовлення. Разом із тим їх використання пов'язане зі значними обчислювальними витратами та потребою у великих навчальних вибірках.

Генеративні дифузійні моделі представляють інший підхід до розв'язання проблеми нечіткого мовлення, зосереджуючись на попередньому очищенні та реконструкції сигналу. Завдяки поетапному видаленню шуму та відновленню спектральної структури вони дозволяють підвищити інформативність ознак, що подаються на вхід системи розпізнавання. Недоліком такого підходу є висока обчислювальна складність і необхідність ретельної оптимізації дифузійного процесу.

Варіаційні автоенкодери застосовуються переважно для моделювання індивідуальних особливостей мовця. Їх використання сприяє персоналізації системи та підвищенню якості розпізнавання для користувачів із нестандартною вимовою або акцентом. Водночас надмірна компресія в просторі може призводити до втрати дрібних фонетичних деталей, що обмежує ефективність цього підходу без достатнього обсягу адаптаційних даних.

Гібридні нейронні системи, які поєднують згорткові та трансформерні компоненти, дозволяють збалансувати локальну чутливість до спектральних ознак і глобальний аналіз контексту. Такі архітектури є придатними для практичних систем розпізнавання мовлення, зокрема в режимі реального часу, проте потребують значних обчислювальних ресурсів і ретельного налаштування.

Таким чином, аналіз сучасних нейронних підходів свідчить, що ефективне розпізнавання нечіткого мовлення досягається або за рахунок масштабних end-to-end моделей, або шляхом інтеграції спеціалізованих модулів попередньої обробки та адаптації. Це обґрунтовує доцільність комбінованих підходів, у яких відновлення мовленнєвого сигналу та персоналізація використовуються для підвищення стійкості систем автоматичного розпізнавання мовлення.

1.4. Теоретичні засади варіаційного навчання (Variational Learning, VAE)

Варіаційне навчання є підходом до побудови та оптимізації імовірнісних моделей, у яких складні або невідомі апостеріорні розподіли, тобто розподіли прихованих змінних за умови спостережуваних даних, оцінюються за допомогою параметризованих сімейств розподілів [31]. Основна ідея цього підходу полягає в заміні прямого обчислення апостеріорного розподілу, яке

для складних моделей є обчислювально-недосяжним, задачею оптимізації, що мінімізує відстань між істинним розподілом і його варіаційною оцінкою [31].

У задачі обробки мовлення варіаційне навчання дозволяє моделювати приховані (латентні) змінні, які відповідають за індивідуальні властивості мовленнєвого сигналу, зокрема тембр, артикуляційні особливості або стиль мовлення. Це є особливо важливим для роботи з нечітким мовленням, де спостережувані дані можуть бути зашумленими або частково спотвореними.

Нехай x - спостережуваний мовленнєвий сигнал, а z – латентна змінна, що описує приховані характеристики мовця або мовлення. Генеративна модель задається спільним розподілом [31]:

$$p_{\theta}(x, z) = p_{\theta}(x|z)p(z) \quad (1.1)$$

де $p(z)$ – розподіл латентних змінних,

$p_{\theta}(x, z)$ – умовний розподіл, що моделює процес генерації мовленнєвого сигналу з латентного простору.

Метою навчання є максимізація правдоподібності спостережуваних даних:

$$\log p_{\theta}(x) = \log \int p_{\theta}(x|z)p(z) dz \quad (1.2)$$

Однак пряме обчислення цього інтеграла є практично неможливим для складних нейронних моделей [31].

Для подолання цієї проблеми вводиться варіаційний розподіл $q_{\phi}(z|x)$, який оцінює істинний апостеріорний розподіл $p_{\theta}(x, z)$. Застосовуючи варіаційний підхід, максимізацію логарифмічної правдоподібності замінюються максимізацією нижньої межі правдоподібності (Evidence Lower Bound, ELBO) [31]:

$$\log p_{\theta}(x) \geq E_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] - KL(q_{\phi}(z|x)||p(z)) \quad (1.3)$$

Перший доданок відповідає якості реконструкції вхідного сигналу, тоді як другий – Kullback-Leibler дивергенції [31], яка регулює латентний простір і змушує варіаційний розподіл бути близьким до апіорного розподілу, тобто заздалегідь заданого імовірнісного розподілу латентних змінних [31].

Таким чином, функція втрат VAE має вигляд:

$$L_{VAE} = -E_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] + KL(q_{\phi}(z|x)||p(z)) \quad (1.4)$$

Латентний простір VAE є компактным і безперервним поданням даних, у якому близькі точки відповідають подібним за властивостями зразкам мовлення. У контексті розпізнавання нечіткого мовлення латентні змінні можуть інтерпретуватися як узагальнені характеристики мовця, що є відносно стабільними навіть за наявності шуму або артикуляційних спотворень.

У системах автоматичного розпізнавання мовлення варіаційні автоенкодери використовуються для:

- персоналізації моделей, шляхом кодування індивідуальних особливостей мовця;
- зменшення впливу індивідуальних особливостей мовців;
- попередньої обробки ознак.

Поєднання VAE з дифузійною моделлю дозволяє розділити задачі відновлення сигналу та адаптації до мовця, що є концептуальною основою запропонованого в даній роботі підходу.

Варіаційні автоенкодери активно використовуються в сучасних інтелектуальних системах завдяки своїй здатності навчатися прихованих закономірностей у даних і узагальнювати їх у компактному вигляді. Вони застосовуються в обробці зображень, аудіо, мовлення, рекомендаційних системах та медичних додатках. Однією з головних переваг VAE є те, що вони дозволяють працювати з неповними, зашумленими або варіативними даними, формуючи стабільні внутрішні представлення. У контексті мовлення це

означає можливість краще враховувати індивідуальні особливості голосу користувача та підвищувати якість роботи систем у складних умовах.

Разом із тим використання VAE має і певні обмеження. Оскільки модель стискає інформацію у компактний внутрішній простір, існує ризик втрати дрібних, але важливих деталей даних.

Висновки до розділу 1

У першому розділі магістерської роботи здійснено комплексний аналіз теоретичних засад автоматичного розпізнавання нечіткого мовлення та сучасного стану розвитку відповідних технологій. Розглянуто фізичну природу мовленнєвого сигналу, особливості його формування та основні акустичні характеристики, що лежать в основі процесів автоматичного аналізу й розпізнавання мовлення. Показано, що мовленнєвий сигнал є складним, багатовимірним об'єктом, чутливим до впливу як зовнішніх акустичних чинників, так і індивідуальних особливостей мовця.

У межах розділу детально проаналізовано поняття нечіткого мовлення та запропоновано його класифікацію за джерелами виникнення, зокрема акустичними, фонетичними, просодичними, лінгвістичними, індивідуальними та технічними чинниками. Показано, що кожен із цих типів нечіткості по-різному впливає на інформативність мовленнєвих ознак і суттєво ускладнює роботу традиційних систем автоматичного розпізнавання мовлення, орієнтованих на «еталонні» умови запису. Наведені статистичні дані підтверджують соціальну значущість проблеми та зростаючу потребу в інклюзивних технологіях.

Огляд сучасних ASR-систем засвідчив, що найбільш високі показники точності досягаються за рахунок використання масштабних нейромережевих моделей, великих обсягів навчальних даних і потужних обчислювальних ресурсів. Разом із тим встановлено, що такі рішення часто є ресурсо-ємними, залежать від хмарної інфраструктури або не забезпечують достатньої стійкості до нечіткої вимови та складних акустичних умов. Це обмежує їх застосування в реальних сценаріях, зокрема в локальних або ресурсно-обмежених системах.

У розділі також розглянуто основні підходи до використання штучних нейронних мереж у розпізнаванні аудіо, зокрема згорткові, рекурентні, трансформерні та гібридні архітектури. Порівняльний аналіз показав, що кожен із підходів має власні переваги та обмеження щодо обробки нечіткого

мовлення, а найбільш перспективними є комбіновані рішення, які поєднують аналіз локальних спектральних ознак із глобальним урахуванням часово-контекстної інформації.

Окрему увагу приділено варіаційному навчанню та варіаційним автоенкодерам як інструментам моделювання прихованих представлень мовця та зменшення впливу індивідуальних мовленнєвих відмінностей.

Таким чином, результати теоретичного аналізу підтверджують, що ефективне розпізнавання нечіткого мовлення потребує поєднання сучасних нейромережевих методів із спеціалізованими механізмами попередньої обробки та адаптації. Це обґрунтовує доцільність подальшого дослідження дифузійних моделей для відновлення мовленнєвого сигналу та їх інтеграції з варіаційним навчанням, що і становить основний напрям подальших розділів даної магістерської роботи.

2. ДИФУЗІЙНА МОДЕЛЬ РОЗПІЗНАВАННЯ НЕЧІТКОГО МОВЛЕННЯ

2.1. Принцип роботи дифузійної моделі та її адаптація для аудіо

Дифузійні моделі належать до класу імовірнісних генеративних моделей, основна ідея яких полягає у поетапному перетворенні складного розподілу даних у простий та зворотному відновленні вихідних даних із цього простого розподілу. На відміну від класичних генеративних підходів, дифузійні моделі реалізують процес генерації як послідовність малих, контрольованих кроків, що забезпечує стабільність навчання та високу якість відновлення даних.

У контексті автоматичного розпізнавання мовлення дифузійна модель використовується не як самостійний генератор, а як модуль попередньої обробки, що виконує реконструкцію зашумленого або нечіткого мовлення перед подачею його на вхід акустичної моделі. Такий підхід дозволяє підвищити інформативність вхідних ознак без зміни архітектури акустичної моделі, що є суттєвою перевагою з погляду сумісності та повторного використання наявних ASR-рішень.

У загальному випадку дифузійна модель складається з двох взаємопов'язаних процесів: прямого (дифузійного) та зворотного (денойзингового). Прямий процес поступово додає шум до даних, перетворюючи їх на випадкову змінну з відомим простим розподілом, тоді як зворотний процес навчається відновлювати дані, крок за кроком усуваючи цей шум.

У контексті обробки мовлення це означає, що чистий мовленнєвий сигнал поступово перетворюється на шум, а нейромережа навчається відновлювати мовлення із спотворених версій.

Процес відновлення сигналу складається з двох етапів:

1. Прямий дифузійний процес визначається як марковський ланцюг, у якому на кожному кроці до сигналу x_0 додається випадковий (гаусівський шум):

$$q(x_t | x_{t-1}) = N(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I) \quad (2.1)$$

де x_0 - початковий (чистий) мовленнєвий сигнал,

$\beta_t \in (0,1)$ – параметр шуму на кроці t ,

x_t – версія сигналу на кроці t після додавання шуму,

I – одинична матриця, що задає гаусівський шум.

Розподіл x_t можна виразити безпосередньо через початковий сигнал x_0 :

$$q(x_t | x_0) = N(x_t; \sqrt{\tilde{a}_t} x_0, (1 - \tilde{a}_t) I) \quad (2.2)$$

де a_t – коефіцієнт збереження частки оригінального сигналу на кроці t ,

$\tilde{a}_t$ – накопичений коефіцієнт деградації сигналу до кроку t .

$$a_t = 1 - \beta_t, \quad \tilde{a}_t = \prod_{s=1}^t a_s \quad (2.3)$$

На будь-якому кроці t ми можемо подати x_t як «спотворену» версію x_0 із доданим гаусівським шумом:

$$x_t = \sqrt{\tilde{a}_t} x_0 + \sqrt{1 - \tilde{a}_t} \varepsilon \quad (2.4)$$

Якщо повторювати цей процес достатню кількість кроків ($t \rightarrow T$), розподіл x_T наближається до чистого шуму, тобто інформація про вихідний сигнал практично втрачається.

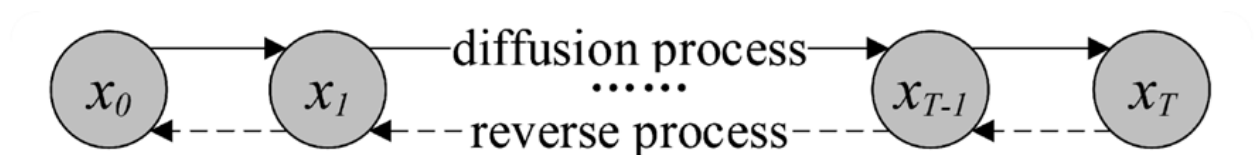


Рис 2.1 – процес дифузії та денойзингу

2. Відновлення сигналу. Під час зворотного процесу модель навчається видаляти шум та поступово відновлювати сигнал:

$$p_{\theta}(x_{t-1} | x_t) = N(x_{t-1}; \mu_{\theta}(x_t, t), \sigma_t^2 I) \quad (2.5)$$

де σ_t^2 – дисперсія шуму на кроці t .

Практично нейромережа навчається оцінювати шум, який було додано на кроці t . Формулу відновлення x_{t-1} можна записати у вигляді:

$$\mu_{\theta}(x_t, t) = \frac{1}{\sqrt{a_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \tilde{a}_t}} \varepsilon_{\theta}(x_t, t) \right) \quad (2.6)$$

де ε_{θ} – оцінка шуму моделлю.

Тоді зворотній хід з кроку t на $t - 1$ має вигляд:

$$x_t = \mu_{\theta}(x_t, t) + \sigma_t z \quad (2.7)$$

де z – додатковий випадковий шум.

Навчання дифузійної моделі здійснюється шляхом мінімізації різниці між передбаченим та реальним шумом:

$$L(\theta) = E_{x_0, t, \varepsilon} \left\| \varepsilon - \varepsilon_{\theta}(\sqrt{\tilde{a}_t} x_0 + \sqrt{1 - \tilde{a}_t} \varepsilon, t) \right\|^2 \quad (2.8)$$

Безпосереднє застосування дифузійних моделей до часових аудіосигналів є обчислювально-складним через високу розмірність даних. Тому в практичних системах дифузійні моделі найчастіше застосовуються до спектральних представлень, зокрема мел-спектрограм:

$$x_0 \in R^{F \times T} \quad (2.9)$$

де F – кількість мел-частотних смуг,

T – кількість часових кадрів.

У цьому випадку дифузійна модель навчається відновлювати структуру спектрограми, пригнічуючи шум і ревербераційні компоненти, але зберігаючи мовленнєві форманти та часову динаміку.

Основними перевагами дифузійної моделі в обробці аудіо є:

- висока якість відновлення зашумлених сигналів;
- стійкість до різних типів шуму та реверберації;

- здатність зберігати інформацію, що потрібна для розпізнавання мовлення.

Водночас до обмежень належать висока обчислювальна складність і значна кількість кроків зворотної дифузії, що потребує оптимізації для застосувань у реальному часі.

Таблиця 2.1 – Основні гіперпараметри та результати навчання дифузійної моделі для розпізнавання нечіткого мовлення

№	Параметр моделі	Позначення	Діапазон / Значення	Вплив на результат
1	Кількість дифузійних кроків	T	250	Значення T = 250 забезпечує компроміс між якістю відновлення та швидкістю інференсу
2	Шумова дисперсія	β_t	від 1e-4 до 0.02 (лінійний розподіл)	Контролює рівень зашумлення; оптимальне зростання покращує стабільність навчання
3	Розмір внутрішнього латентного представлення	d_latent	256	Баланс між деталізацією спектра та стабільністю
4	Тип нейронної архітектури	–	U-Net + Self-Attention	Забезпечує багатомасштабне узгодження часових залежностей
5	Тип вхідного подання	–	Мел-спектрограма 80 смуг	Зменшує вимоги до пам'яті й зберігає перцептивну якість
6	Оптимізатор	–	AdamW (lr = 1e-4, $\beta_1=0.9$, $\beta_2=0.999$)	Сприяє стабільній збіжності на великому датасеті
7	Розмір батчу	B	16	Впливає на узгодження статистик нормалізації; оптимальне B забезпечує рівновагу

				між стабільністю і швидкістю
8	Функція втрат	L	MSE + перцептивна STFT-втрата	Забезпечує одночасне мінімізування шуму і покращення розбірливості

У таблиці 2.1 наведено основні гіперпараметри дифузійної моделі, використаної для попередньої обробки нечіткого мовлення, а також їх вплив на якість навчання та кінцевий результат розпізнавання. Вибір параметрів здійснювався з урахуванням компромісу між точністю відновлення мовленнєвого сигналу, стабільністю навчання та обчислювальною ефективністю.

Кількість дифузійних кроків $T = 250$ була обрана як оптимальна з точки зору балансу між якістю реконструкції та швидкістю інференсу, тобто етапу роботи навченої моделі, на якому виконується відновлення мовлення для нових входних даних без оновлення її параметрів. Менша кількість кроків дозволяє зменшити час обробки одного мовленнєвого фрагмента, водночас зберігаючи достатній рівень деталізації відновленої спектральної структури. Практичні експерименти показали, що подальше збільшення T не дає пропорційного покращення якості, але суттєво підвищує обчислювальні витрати.

Шумова дисперсія β_t змінюється в діапазоні від 10^{-4} до 0.02 за лінійним графіком, що забезпечує поступове і контрольоване зашумлення сигналу в прямому дифузійному процесі. Такий вибір запобігає деградації корисних мовленнєвих ознак на ранніх етапах дифузії.

Розмір внутрішнього латентного представлення моделі становить 256, що забезпечує достатню ємність для кодування спектральних і часових характеристик мовлення без надмірного ускладнення моделі. Це значення дозволяє зберігати баланс між деталізацією мел-спектрограм та стабільністю навчального процесу.

Як базову нейронну архітектуру використано модель типу U-Net із вбудованими механізмами самоуваги, що забезпечує багато-масштабне узгодження локальних і глобальних часових залежностей у спектральному представленні мовлення. Така архітектура є особливо ефективною для задач відновлення зашумлених і спотворених аудіосигналів.

Для оптимізації параметрів моделі використано оптимізатор AdamW з коефіцієнтом навчання $lr = 10^{-4}$, що сприяє стабільній та передбачуваній збіжності на великому навчальному датасеті. Розмір батчу $B = 16$ було обрано як компроміс між швидкістю навчання та стабільністю статистик нормалізації.

2.2. Архітектура моделі

Архітектура дифузійної моделі, використаної в даній роботі, спроектована з урахуванням специфіки мовленнєвих даних та вимог систем автоматичного розпізнавання нечіткого мовлення.

Вхідними даними для дифузійної моделі є мел-спектрограми мовленнєвого сигналу, яка являє собою двовимірне часово-частотне представлення з фіксованою кількістю часових смуг.

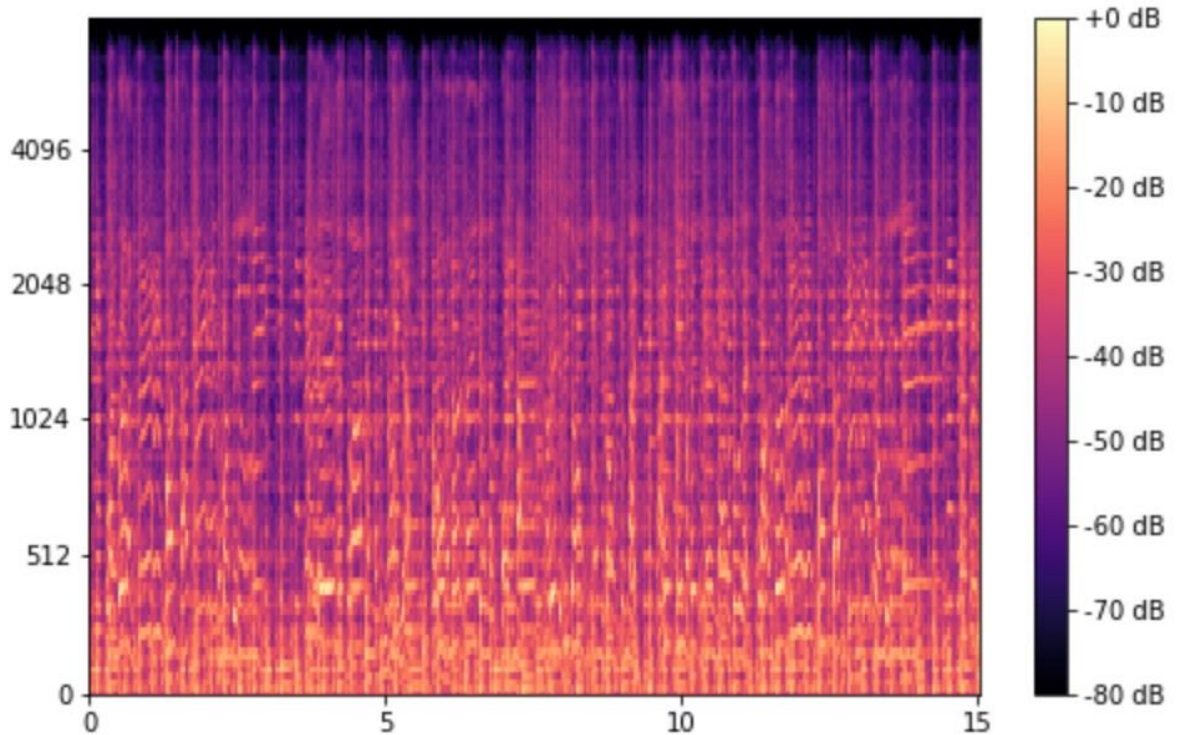


Рис 2.2 – Приклад мел-спектрограми мовленнєвого сигналу

На рисунку 2.2 зображено мел-спектрограму мовленнєвого сигналу, яка є двовимірним часово-частотним поданням аудіоданих. По осі абсцис відкладено час у секундах, що відображає часову динаміку мовлення. По осі ординат відкладено частотні складові сигналу, де числові значення (512, 1024, 2048, 4096 Hz) відповідають рівням частот і вказують, до яких частотних діапазонів належить енергія мовленнєвого сигналу. Кольорова шкала відображає інтенсивність сигналу в децибелах, де світліші відтінки відповідають вищій енергії, а темніші – нижчій.

Мел-спектрограма була обрана як основне вхідне подання для дифузійної моделі з огляду на її здатність ефективно відображати акустичні властивості мовлення у формі, близькій до особливостей слухового сприйняття людини.

Мел-шкала реалізує нелінійне перетворення частотної осі, у якому нижні частоти представлені з вищою роздільною здатністю, ніж високі. Саме в цьому діапазоні зосереджена основна мовленнєва інформація, зокрема формантна структура та переходи між фонемами, що є критично важливими для автоматичного розпізнавання мовлення.

Додатковою перевагою мел-спектрограм є зменшення розмірності даних порівняно з повним спектром, що дозволяє знизити обчислювальні витрати без суттєвої втрати інформативності. Це є особливо важливим у контексті дифузійної моделі, яка виконує багатоетапну обробку сигналу.

Крім того, мел-спектрограма є стандартним і добре дослідженим поданням у сучасних ASR-системах, що забезпечує сумісність дифузійного модуля з існуючими архітектурами акустичних моделей. Вони демонструють відносну стійкість до шуму та реверберації, оскільки підкреслюють значущі компоненти мовлення. Це робить мел-спектрограму доцільним вибором для системи розпізнавання нечіткого мовлення.

В основі архітектури лежить модифікована мережа типу U-Net, яка складається з енкодерної та декодерної частин. З метою покращення врахування довготривалих часових залежностей у мовленні до архітектури включено блоки self-attention, розміщені на проміжних рівнях U-Net. Механізм самоуваги дозволяє моделі встановлювати взаємозв'язки між віддаленими

часовими сегментами спектрограми, що є особливо важливим у випадках нечіткої артикуляції або сильних акустичних спотворень.

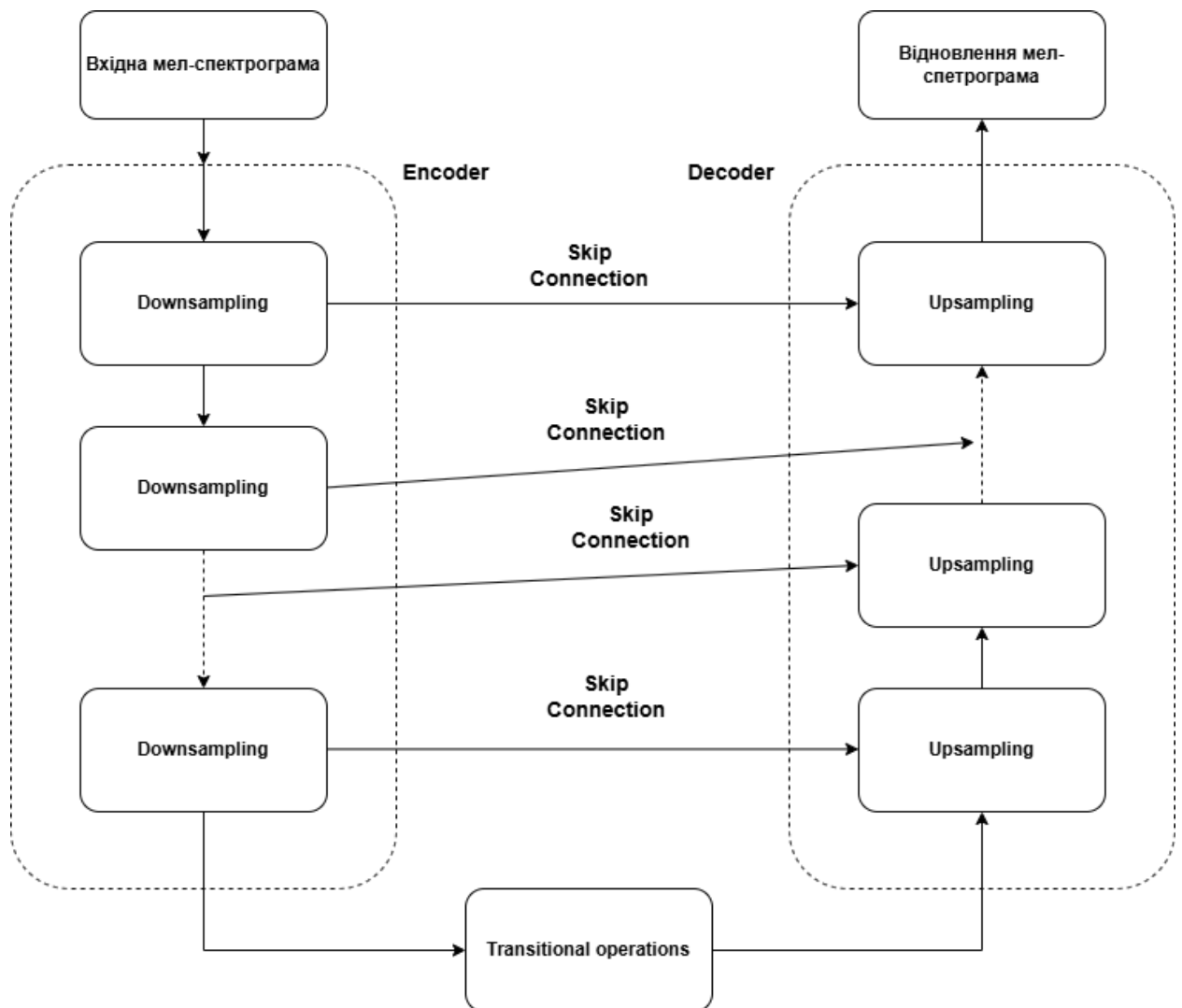


Рис. 2.3 – Архітектура дифузійної моделі на основі U-Net для відновлення мел-спектрограми мовлення

На рисунку 2.3 представлено U-Net подібну архітектуру, яка використовується в дифузійній моделі для покрокового очищення та відновлення мел-спектрограми. Основна ідея U-Net полягає в тому, що модель одночасно вчиться виділяти узагальнені ознаки мовлення і зберігати дрібні деталі через skip-з'єднання між симетричними рівнями.

Ліва частина архітектури відповідає енкодерному етапу і складається з послідовних блоків зменшення роздільної здатності (Downsampling). На

кожному рівні виконується стискання часово-частотного представлення з одночасним формуванням більш узагальнених ознак, що дозволяє моделі виявляти глобальну структуру мовлення та відокремлювати її від шумових компонентів.

У центральній частині архітектури розміщено перехідний рівень, у якому зосереджено найбільш стиснуте та інформативне подання сигналу. На цьому етапі модель узагальнює основні спектральні та часові ознаки мовлення, які використовуються як основа для подальшого відновлення сигналу.

Права частина схеми реалізує декодерний етап і складається з блоків збільшення роздільної здатності (Upsampling), які поступово відновлюють детальну структуру мел-спектрограми. Під час цього процесу модель реконструює часово-частотні компоненти мовлення, зменшуючи вплив шуму та реверберації.

Важливу роль у роботі архітектури відіграють пропуски з'єднання між відповідними рівнями енкодера та декодера. Вони забезпечують передачу інформативних ознак безпосередньо до етапу відновлення, що дозволяє зберігати чіткі спектральні деталі, пов'язані з формантною структурою та переходами між фонемами. Завдяки цьому відновлена мел-спектрограма є більш розбірливою та придатною для подальшого автоматичного розпізнавання мовлення.

На виході дифузійної моделі формується очищене спектральне подання мовлення, яке використовується як вхід для акустичної моделі системи автоматичного розпізнавання мовлення.

Внутрішній латентний простір моделі має розмірність 256 і використовується для кодування узагальнених спектральних і часових характеристик мовлення. Це представлення є компактним, але достатньо інформативним для відновлення ключових мовленнєвих ознак. Такий вибір розмірності забезпечує баланс між деталізацією та стабільністю навчання, запобігаючи перенавчанню та надмірному ускладненню моделі.

На виході дифузійна модель формує відновлену мел-спектрограму або оцінку шумової компоненти, залежно від обраної формалізації задачі навчання. Отримане спектральне представлення використовується як вхід для акустичної моделі системи автоматичного розпізнавання мовлення без необхідності змінювати її архітектуру.

Таким чином, дифузійна модель виконує роль фронтенд-модуля попередньої обробки, який підвищує інформативність вхідних ознак і забезпечує кращу стійкість системи до нечіткого мовлення та складних акустичних умов.

2.3. Використання варіаційного автоенкодера

Варіаційний автоенкодер (Variational Autoencoders, VAE) є класом генеративної нейронної моделі, що поєднує автоенкодерну архітектуру з імовірнісним підходом до моделювання даних. На відміну від класичних автоенкодерів, які виконують детерміноване стискання та відновлення вхідних даних, VAE формують імовірнісне латентне представлення, що дозволяє моделі узагальнювати приховані закономірності у даних і працювати зі складною варіативністю мовленнєвих сигналів [31].

У контексті автоматичного розпізнавання мовлення варіаційні автоенкодери використовуються для моделювання прихованих характеристик мовця та мовлення, які не спостерігаються безпосередньо у вхідному сигналі, але істотно впливають на його акустичну реалізацію. До таких характеристик належать тембр голосу, темп мовлення, індивідуальні артикуляційні особливості та стабільні фонетичні відхилення. Формування компактного латентного представлення дозволяє відокремити ці фактори від шумових і випадкових спотворень.

Однією з ключових переваг використання VAE є здатність зменшувати вплив індивідуальних відмінностей між мовцями шляхом проєкції мовленнєвих ознак у структурований латентний простір. Таке узагальнене

представлення є менш чутливим до змін акустичних умов і сприяє підвищенню стабільності подальшого розпізнавання. Це особливо важливо для нечіткого мовлення, де традиційні подання можуть значно деградувати під впливом шуму, реверберації або нестандартної вимови.

У практичних ASR-системах варіаційні автоенкодери можуть використовуватися як окремий модуль адаптації або персоналізації. Зокрема, латентний вектор, сформований енкодером VAE, може слугувати додатковою умовною інформацією для акустичної моделі, дозволяючи їй адаптувати процес декодування до конкретного користувача без повного перенавчання всієї системи. Такий підхід знижує вимоги до обсягу персональних даних і забезпечує гнучкість системи.

Крім того, VAE можуть застосовуватися для попередньої обробки спектральних ознак мовлення. Завдяки імовірнісному характеру навчання модель навчається відновлювати найбільш характерні та стабільні компоненти сигналу, пригнічуючи випадкові або шумові спотворення. Це робить варіаційні автоенкодери корисними для роботи з нечітким мовленням, де важливо зберегти лінгвістично значущу інформацію за умов обмеженої якості сигналу.

У межах даної роботи варіаційні автоенкодери розглядаються як складова комбінованого підходу до обробки нечіткого мовлення. Їх використання у поєднанні з дифузійною моделлю дозволяє розділити задачі відновлення мовленнєвого сигналу та адаптації до індивідуальних особливостей мовця. Такий підхід забезпечує підвищення стійкості системи автоматичного розпізнавання мовлення до шуму, реверберації та нечіткої артикуляції.

Як показано у розділі 1, варіаційний автоенкодер моделює приховані змінні мовленнєвого сигналу шляхом оцінки апостеріорного розподілу латентних змінних відповідно до формул (1.1) – (1.4). Отримане латентне представлення дозволяє узагальнити індивідуальні особливості мовця та зменшити вплив випадкових акустичних варіацій.

У свою чергу, дифузійна модель, формально описана рівняннями (2.1)–(2.6), реалізує процес поступового зашумлення мовленнєвого сигналу та його подальшого покрокового відновлення. Ключовою особливістю запропонованого підходу є використання латентного представлення, сформованого VAE, як умовної інформації для зворотного дифузійного процесу.

Нехай z – латентний вектор, отриманий з енкодера VAE відповідно до апостеріорного розподілу. У запропонованій архітектурі цей вектор використовується для модифікації зворотного дифузійного переходу. Зокрема, умовний зворотний перехід визначається як:

$$p_{\theta}(x_{t-1} | x_t, z) \quad (2.10)$$

що є розширенням стандартного зворотного процесу.

Відповідно, нейронна мережа дифузійної моделі навчається оцінювати шумову компоненту з урахуванням латентної змінної [31]:

$$\widehat{\varepsilon}_{\theta} = \varepsilon_{\theta}(x_t, t, z) \quad (2.11)$$

де x_t – зашумлене спектральне подання на кроці t ,

z – містить узагальнену інформацію про мовця.

Такий підхід дозволяє коригувати процес очищення мовлення залежно від індивідуальних особливостей вимови, тембру та артикуляції.

Навчання дифузійної частини моделі з інтеграцією VAE здійснюється шляхом мінімізації модифікованої функції втрат:

$$L_{diff} = E_{x_0, t, \varepsilon, z} \|\varepsilon - \varepsilon_{\theta}(x_t, t, z)\|^2 \quad (2.12)$$

Для узгодженого навчання всієї системи використовується сумарна функція втрат [31]:

$$L_{total} = L_{diff} + L_{VAE} \quad (2.12)$$

З практичної точки зору така інтеграція означає, що дифузійна модель не лише відновлює мовленнєвий сигнал із шуму, а робить це з урахуванням стабільних мовленнєвих характеристик, закодованих у латентному просторі VAE. Це зменшує варіативність між різними мовцями, підвищує узгодженість спектральних ознак і сприяє формуванню більш інформативного вхідного подання для акустичної моделі системи автоматичного розпізнавання мовлення.

Таким чином, варіаційний автоенкодер у запропонованій системі виконує роль механізму адаптації та персоналізації, тоді як дифузійна модель відповідає за поетапне очищення та реконструкцію мовленнєвого сигналу. Їх поєднання утворює єдину імовірнісну архітектуру, орієнтовану на ефективну обробку нечіткого мовлення в реальних акустичних умовах.

2.4. Алгоритм розпізнавання нечіткого мовлення на основі запропонованої моделі

Запропонований алгоритм автоматичного розпізнавання нечіткого мовлення базується на поєднанні дифузійної моделі попередньої обробки, варіаційного автоенкодера для моделювання індивідуальних особливостей мовця та акустичної моделі системи ASR. Така багатокомпонентна структура дозволяє розділити етапи очищення мовленнєвого сигналу, адаптації до мовця та безпосереднього розпізнавання, що підвищує загальну стійкості системи в складних акустичних умовах до шуму, реверберації та нечіткої артикуляції.

Перед застосуванням дифузійної моделі мовленнєвий сигнал необхідно перетворити у придатне для нейромережевої обробки представлення. Для

цього з часової форми аудіосигналу формується мел-спектрограма, яка стисло відображає його спектрально-часову структуру та зберігає значущі характеристики мовлення. Отримане подання використовується як вхідний формат для дифузійної моделі попередньої обробки та подальшого автоматичного розпізнавання мовлення.

Отримання мел-спектрограми поділяється на декілька етапів:

1. На першому етапі сигнал підсилюють у високочастотній області [10], оскільки саме цей діапазон є важливим для розпізнавання мовних ознак та сприймається людським слухом найбільш виразно:

$$y[n] = x[n] - ax[x - 1] \quad (2.13)$$

де a – коефіцієнт підсилення високих частот,
 $x[n]$ – вхідний мовленнєвий сигнал у часовій області,
 $y[n]$ – сигнал після попереднього підсилення.

2. Далі сигнал $y[n]$ ділять на перекривні фрейми: тривалість ≈ 20 – 25 мс, крок ≈ 10 мс.

3. Перед спектральним аналізом до кожного фрейма застосовується віконна функція, щоб згладити різкі переходи між сусідніми фреймами. Найчастіше використовується вікно Гемінга [10]:

$$H[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad (2.14)$$

$$\tilde{x}_m[n] = y_m[n] \cdot H[n], \quad (2.15)$$

де $n = 0, 1, \dots, N - 1$,
 m – номер фрейму в часовій області,
 $H[n]$ – віконна функція,
 n – індекс відліку всередині фрейму.

4. Наступним кроком є короткочасне перетворення Фур'є (STFT) [10], яке дозволяє перейти від часового представлення до частотного:

$$X_m[k] = \sum_{n=0}^{N-1} \tilde{x}_m[n] \cdot e^{-\frac{j2\pi kn}{N}} \quad (2.16)$$

де $k = 0, 1, \dots, N - 1$,

N – довжина фрейму,

$X_m[k]$ – спектр сигналу у частотній області.

5. Потім обчислюється спектр потужності, який характеризує розподіл енергії мовленнєвого сигналу за частотами [10].

$$P_m[k] = |X_m[k]|^2 \quad (2.17)$$

6. Для врахування особливостей слухового сприйняття людини спектр потужності проектується на мел-шкалу, яка є нелінійним перетворенням частотної осі [10]:

$$f_{mel} = 2595 \cdot \log_{10} \left(1 + \frac{f}{700} \right) \quad (2.18)$$

7. Для кожного фільтра на мел-шкалі обчислюють енергію сигналу:

$$S_m[k] = \sum_k P_m[k] \cdot H_m[k] \quad (2.19)$$

де $H_m[k]$ – частотна характеристика m -го фільтра.

8. До отриманих мел-енергій застосовується логарифмічне перетворення [10]:

$$MelSpec_m[k] = \log(S_m[k] + \epsilon) \quad (2.19)$$

де ϵ – мала додатна константа, що запобігає логарифмуванню нуля [10].



Рис 2.4 – Алгоритм отримання мел-спектрограми

У результаті формується мел-спектрограма. Це подання стисло та інформативно описує спектрально-часову структуру мовлення і використовується як вхідне подання для дифузійної моделі попередньої обробки та подальшого автоматичного розпізнавання мовлення.



Рис. 2.5 – Алгоритм розпізнавання мовлення в розробленій системі

Далі сформована мел-спектрограма подається на вхід варіаційного автоенкодера. Енкодер VAE виконує проєкцію спектрального подання у латентний простір, формуючи компактний вектор, що узагальнює стабільні індивідуальні характеристики мовця, такі як тембр голосу, типова артикуляція та спектральні особливості. Отримане латентне представлення використовується як умовна інформація на наступному етапі, що дозволяє адаптувати процес очищення мовлення до конкретного користувача.

Наступним етапом є дифузійна попередня обробка мовлення. На цьому кроці мел-спектрограма, яка містить спотворення, проходить зворотний дифузійний процес, у межах якого модель поетапно видаляє шумову

компоненту. На кожному кроці зворотної дифузії нейронна мережа оцінює шум із урахуванням як поточного рівня зашумлення, так і латентного вектора, отриманого з VAE. Це дозволяє більш точно відновлювати спектральну структуру мовлення, зменшуючи вплив нечіткої вимови та акустичних перешкод.

Після завершення всіх кроків зворотного дифузійного процесу формується очищена мел-спектрограма, яка характеризується зменшеним рівнем шуму та підвищеною розбірливістю мовленнєвих ознак. Отримане спектральне подання безпосередньо передається на вхід акустичної моделі системи автоматичного розпізнавання мовлення без необхідності змінювати її архітектуру або параметри.

Акустична модель виконує перетворення очищених спектральних ознак у послідовність ймовірностей фонем, після чого застосовується декодувальний алгоритм із використанням мовної моделі та словника відповідної мови. На цьому етапі відбувається формування фінальної текстової транскрипції мовленнєвого сигналу.

Таким чином, запропонований алгоритм реалізує послідовний ланцюг обробки, у якому кожен етап виконує чітко визначену функцію: попередня спектральна обробка забезпечує компактне та інформативне подання сигналу, варіаційний автоенкодер формує латентне представлення мовця, дифузійна модель здійснює поетапне очищення мовлення, а акустична модель ASR виконує безпосереднє розпізнавання. Таке модульне поєднання дозволяє підвищити точність розпізнавання нечіткого мовлення та забезпечує гнучкість системи щодо подальшого розширення або адаптації до нових умов.

2.5. Оптимізація гіперпараметрів та процес навчання моделі

Ефективність системи автоматичного розпізнавання нечіткого мовлення на основі дифузійної моделі значною мірою визначається правильним вибором гіперпараметрів і організацією процесу навчання. Оскільки

запропонована система поєднує модуль попередньої обробки аудіосигналу на основі дифузійної моделі та акустичну модель типу Conformer-Transducer, оптимізація виконується з урахуванням особливостей обох компонентів.

Процес оптимізації спрямований на досягнення компромісу між трьома ключовими характеристиками системи:

- точністю розпізнавання мовлення, що оцінюється за метриками WER та CER;
- обчислювальною складністю;
- стабільністю навчання.

На відміну від класичних ASR-систем, у запропонованому підході оптимізація охоплює не лише параметри акустичної моделі, а й гіперпараметри дифузійного процесу, які безпосередньо впливають на якість реконструкції мовленнєвого сигналу.

Ключовими гіперпараметрами дифузійного модуля є:

- кількість дифузійних кроків T ;
- розклад коефіцієнтів шуму β_t ;
- оптимізатор AdamW;
- розмір батчу.

Дифузійна модель, що використовується як модуль попередньої обробки нечіткого мовлення, містить значну кількість гіперпараметрів, які безпосередньо впливають на якість реконструкції мовленнєвого сигналу, стабільність навчання та обчислювальну складність системи. Некоректний вибір цих параметрів може призводити як до недостатнього очищення сигналу, так і до надмірної затримки обробки, що є критичним для систем реального часу. Тому оптимізація гіперпараметрів дифузійної моделі є одним із ключових етапів розробки запропонованої системи.

Одним із основних гіперпараметрів дифузійної моделі є кількість кроків дифузії T , яка визначає глибину процесу зашумлення та подальшого відновлення сигналу. У класичних дифузійних моделях збільшення T

призводить до більш плавного та точного відновлення даних, оскільки кожен крок зворотного процесу видаляє лише невелику частину шуму. Проте в задачі попередньої обробки мовлення це супроводжується суттєвим зростанням часу обробки.

У межах даної роботи було проведено серію експериментів з різними значеннями T у діапазоні від 25 до 400. Результати показали, що при $T < 30$ якість реконструкції є недостатньою: у відновленому сигналі зберігаються залишкові шуми та спотворення артикуляційних ознак. Водночас значення $T > 300$ не дають суттєвого приросту якості розпізнавання, але призводять до неприйнятної затримки обробки.

Важливим гіперпараметром дифузійної моделі є розклад коефіцієнтів шуму β_t який визначає інтенсивність спотворення мовленнєвого сигналу на кожному кроці прямого дифузійного процесу. Від вибору цього розкладу залежить характер втрати інформації у процесі зашумлення, стабільність навчання зворотного процесу та якість подальшої реконструкції сигналу.

У межах даної роботи було обрано лінійний розклад коефіцієнтів шуму. Лінійний розклад характеризується поступовим і контрольованим зростанням рівня шуму на кожному кроці дифузії, що є важливим для задачі попередньої обробки мовленнєвих сигналів. Такий підхід забезпечує передбачувану деградацію сигналу в прямому процесі та спрощує навчання зворотної моделі, оскільки зміни між сусідніми кроками мають однорідний характер.

Вибір лінійного розкладу зумовлений кількома чинниками. По-перше, він є обчислювально простим і стабільним у реалізації, що має принципове значення для систем реального часу. По-друге, лінійний розклад дозволяє зберігати важливі спектральні характеристики мовлення на ранніх етапах дифузії, запобігаючи їх передчасному руйнуванню. По-третє, експериментальні результати показали, що при фіксованій кількості дифузійних кроків лінійний розклад забезпечує стабільну якість реконструкції сигналу та відтворювані результати під час навчання.

Розмір батчу є одним із ключових гіперпараметрів процесу навчання дифузійної моделі, який безпосередньо впливає на стабільність оптимізації, швидкість збіжності та узагальнювальну здатність моделі.

Батч — це підмножина навчальних даних, яка використовується для обчислення функції втрат і виконання одного кроку оновлення параметрів моделі в процесі навчання нейронної мережі. У задачах обробки мовленнєвих сигналів вибір розміру батчу ускладнюється високою розмірністю вхідних даних та значними вимогами до обчислювальних ресурсів, особливо при використанні дифузійних моделей.

З одного боку, збільшення розміру батчу дозволяє зменшити дисперсію оцінки градієнта та забезпечує більш стабільні оновлення параметрів моделі. Це може позитивно впливати на швидкість збіжності та відтворюваність результатів. З іншого боку, великі батчі потребують значних обсягів відеопам'яті та можуть призводити до зниження здатності моделі до узагальнення, що є небажаним у задачах розпізнавання нечіткого мовлення з високою варіативністю акустичних умов.

У межах даної роботи було обрано розмір батчу, рівний 16, який забезпечує збалансований компроміс між стабільністю оцінки градієнтів і допустимим споживанням обчислювальних ресурсів. Застосування такого розміру батчу дозволяє зменшити дисперсію градієнтів порівняно з навчанням на дуже малих батчах, водночас не призводячи до надмірного навантаження на відеопам'ять, що є критичним при навчанні дифузійних моделей.

Під дисперсією градієнтів у контексті навчання нейронних мереж розуміють міру розкиду значень оціненого градієнта функції втрат відносно його математичного сподівання, яка виникає внаслідок використання обмеженої підмножини навчальних даних під час одного кроку оптимізації. Оскільки градієнт обчислюється не на всій вибірці, а лише на батчі скінченного розміру, його значення є стохастичною оцінкою істинного градієнта.

Під час навчання нейронної мережі необхідно знайти такі значення параметрів (ваг), за яких помилка моделі на навчальних даних буде мінімальною. Для цього використовується алгоритм оптимізації, який визначає, у якому напрямку і на яку величину змінювати параметри моделі після кожного кроку навчання. У даній роботі для навчання дифузійної та акустичної моделей застосовано оптимізатор AdamW, який є одним із найбільш ефективних і стабільних методів оптимізації глибоких нейронних мереж.

У процесі навчання нейронної мережі після обчислення функції втрат визначається її градієнт — вектор, який показує, як саме слід змінювати параметри моделі, щоб зменшити помилку. Класичний стохастичний градієнтний спуск змінює параметри у напрямку протилежному градієнту, однак такий підхід має суттєві недоліки: він чутливий до вибору швидкості навчання та нестабільний у випадку шумних або нерівномірних градієнтів, що є типовим для задач обробки мовлення.

Оптимізатор Adam був розроблений для подолання цих обмежень шляхом використання адаптивної швидкості навчання, яка автоматично підлаштовується під кожен параметр моделі окремо. AdamW є подальшим удосконаленням цього підходу.

Висновки до розділу 2

У другому розділі магістерської роботи розроблено та детально досліджено дифузійну модель попередньої обробки нечіткого мовлення, орієнтовану на підвищення ефективності систем автоматичного розпізнавання мовлення в умовах акустичних і артикуляційних спотворень. Основна увага розділу зосереджена на теоретичному обґрунтуванні вибраного підходу.

Розглянуто принципи роботи дифузійних моделей як імовірнісних генеративних методів, що реалізують процес відновлення даних шляхом поетапного усунення шуму. Показано, що адаптація дифузійного підходу до задач обробки мовлення є доцільною, оскільки мовленнєвий сигнал природно піддається деградації в реальних умовах, а його відновлення може бути формалізоване як послідовний процес денойзингу. Обґрунтовано використання спектральних представлень, зокрема мел-спектрограм, як вхідних даних для дифузійної моделі з метою зменшення обчислювальної складності та збереження значущих характеристик мовлення.

Описано роботу запропонованої дифузійної моделі, яка базується на U-подібній згортковій мережі з механізмами самоуваги. Таке поєднання дозволяє ефективно враховувати як локальні часово-частотні структури мовлення, так і глобальні контекстні залежності, що є критично важливим для коректного відновлення нечіткого або зашумленого сигналу. Вибір U-подібної архітектури обґрунтовано з точки зору її здатності до багато-масштабної реконструкції спектральної структури мовлення.

Розглянуто інтеграцію варіаційного автоенкодера у загальну структуру моделі. Показано, що використання варіаційного навчання дозволяє сформувати компактний латентний простір, у якому моделюються індивідуальні особливості мовлення. Це створює передумови для персоналізації системи та підвищення її стійкості до варіативності між мовцями, що є особливо важливим у задачі розпізнавання нечіткого мовлення.

Запропоновані алгоритмічні рішення створюють передумови для підвищення точності автоматичного розпізнавання мовлення в складних акустичних умовах і можуть бути інтегровані в існуючі ASR-системи без суттєвих змін їх внутрішньої структури. Отримані результати є базою для практичної реалізації та експериментальної перевірки ефективності запропонованого підходу, що здійснюється у наступному розділі магістерської роботи.

3. РЕАЛІЗАЦІЯ СИСТЕМИ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВЛЕННЯ

3.1. Вибір інструментів і середовища реалізації (Python, PyTorch, Librosa тощо)

Для реалізації системи розпізнавання нечіткого мовлення було обрано мову програмування Python, оскільки вона забезпечує найбільш гнучке середовище для розробки моделей машинного навчання та обробки аудіосигналів. Python має широку підтримку наукових бібліотек, що дозволяє поєднувати алгоритмічну частину з математичними обчисленнями, візуалізацією, аналізом і тестуванням моделей. Його відкритий код і масштабована екосистема роблять можливим створення систем будь-якого рівня складності – від прототипів до промислових застосувань.

У контексті дифузійних моделей Python забезпечує простоту реалізації обчислювальних процесів, адаптацію до різних графічних прискорювачів і підтримку паралельних обчислень, що істотно скорочує час навчання системи [29].

Основним фреймворком для побудови системи автоматичного розпізнавання мовлення став PyTorch, що зумовлено його гнучкістю та зручністю реалізації складних нейромережевих архітектур. Він підтримує ефективне навчання моделей із використанням апаратного прискорення на GPU, а також має інструменти для обробки аудіосигналів, зокрема бібліотеку TorchAudio, що спрощує реалізацію перетворення мовлення від сигналу до тексту.

Для попередньої обробки аудіосигналів використовується бібліотека Librosa, яка надає набір інструментів для аналізу, перетворення та підготовки мовленнєвого сигналу у цифровому вигляді. За допомогою Librosa виконуються базові етапи обробки аудіо, зокрема зчитування сигналу,

нормалізація амплітуди, поділення сигналу на фрейми, а також обчислення спектрограми.

```
def load_audio(file_obj_or_bytes=None, target_sr=16000, duration_sec=3.0): 3 usages
    if HAVE_LIBROSA and file_obj_or_bytes is not None:
        try:
            if hasattr(file_obj_or_bytes, "read"):
                data = file_obj_or_bytes.read()
            else:
                data = file_obj_or_bytes
            y, sr = librosa.load(io.BytesIO(data), sr=target_sr, mono=True)
            return y.astype(np.float32), int(sr)
        except Exception:
            pass

    sr = int(target_sr)
    t = np.linspace(0, duration_sec, int(duration_sec * sr), endpoint=False)
    y = 0.3 * np.sin(2 * np.pi * 220 * t) + 0.05 * np.random.randn(len(t))
    return y.astype(np.float32), sr
```

Рис 3.1. Фрагмент коду завантаження аудіо до системи

У процесі реалізації системи автоматичного розпізнавання мовлення також використано бібліотеки NumPy, SciPy, Pandas та Matplotlib, які надають інструменти для чисельних розрахунків, обробки даних і аналізу результатів.

```
def plot_spectrogram(S, freqs, times, title): 2 usages
    fig, ax = plt.subplots(figsize=(6, 3))
    im = ax.imshow(S, aspect="auto", origin="lower",
                   extent=[float(times.min()), float(times.max()),
                           float(freqs.min()), float(freqs.max())])
    ax.set_title(title, fontsize=11)
    ax.set_xlabel("Час, с")
    ax.set_ylabel("Частота, Гц")
    fig.colorbar(im, ax=ax, shrink=0.8)
    fig.tight_layout()
    return fig
```

Рис 3.2. Фрагменту коду, що відповідає за візуалізацію спектрограми за допомогою бібліотеки Matplotlib

```

def compute_spectrogram(y, sr, n_fft=1024, hop=256, use_mel=True): 2 usages
    if HAVE_LIBROSA and use_mel:
        S_pow = librosa.feature.melspectrogram(y=y, sr=sr, n_fft=n_fft,
                                                hop_length=hop, n_mels=128, power=2.0)
        S_db = librosa.power_to_db(S_pow, ref=np.max)
        freqs = np.linspace(0, sr/2, S_db.shape[0])
        times = np.arange(S_db.shape[1]) * hop / sr
        return S_db, freqs, times, True

    win = np.hanning(n_fft).astype(np.float32)
    frames = max(1, 1 + (len(y) - n_fft) // hop)
    spec = []
    for i in range(frames):
        seg = y[i * hop: i * hop + n_fft]
        if len(seg) < n_fft:
            seg = np.pad(seg, (0, n_fft - len(seg)))
        seg = seg * win
        S = np.abs(np.fft.rfft(seg)) + 1e-8
        spec.append(S)
    S_arr = np.array(spec).T
    freqs = np.linspace(0, sr/2, S_arr.shape[0])
    times = np.arange(S_arr.shape[1]) * hop / sr
    S_log = 20.0 * np.log10(S_arr / np.max(S_arr))
    return S_log, freqs, times, True

```

Рис. 3.3. Фрагмент коду, що відповідає за формування спектрограми за допомогою бібліотеки NumPy

Для роботи з аудіофайлами у системі використовуються фреймворк SoundFile, який забезпечує надійне зчитування та запис аудіоданих без втрати якості.

Для управління процесом тренування нейромережі застосовано бібліотеку PyTorch Lightning, яка забезпечує організацію навчального циклу та відокремлення логіки моделі від процедур навчання й оцінювання. PyTorch Lightning спрощує реалізацію етапів тренування та тестування.

Розробку, тестування та візуалізацію системи здійснено у середовищі JupyterLab, що забезпечує інтерактивну роботу з кодом та результатами обчислень в єдиному робочому просторі. JupyterLab застосовано для перевірки етапів попередньої обробки аудіосигналу, аналізу результатів навчання нейромережі та відображення результатів роботи системи.

Для реалізації користувацького інтерфейсу розробленої системи автоматично розпізнавання нечіткого мовлення використано фреймворк StreamLit, який дозволяє створювати інтерактивні застосунки. За допомогою

StreamLit реалізовано інтерфейс завантаження аудіофайлів, запису голосу, запуску процесу попередньої обробки та розпізнавання мовлення, а також візуалізації результатів у зручному для користувача вигляді.

```
def run_streamlit(): 1 usage
    if not HAVE_STREAMLIT:
        print("Не знайдено streamlit. Встановіть: pip install streamlit")
        sys.exit(1)

    st.set_page_config(page_title="Розпізнавання нечіткого мовлення", page_icon="🔊", layout="wide")
    st.title("🔊 Розпізнавання нечіткого мовлення – демо інтерфейс (Streamlit)")
    st.caption("Завантаження аудіо → дифузійне очищення (демо) → спектрограми → текстовий результат")

    with st.sidebar:
        st.header("Параметри")
        steps = st.slider("К-сть кроків очищення (умовно)", 4, 50, 20, 2)
        denoise_strength = st.slider("Інтенсивність очищення", 0.0, 1.0, 0.35, 0.05)
        sr_sel = st.selectbox("Частота дискретизації", [16000, 22050, 24000], index=0)
        st.markdown("---")
        st.info("За відсутності залежностей демо працює на синтетичному сигналі.")

    upl = st.file_uploader("Завантажте аудіо (WAV/MP3/FLAC)", type=["wav", "mp3", "flac"])

    if st.button("Обробити") or upl is not None:
        with st.spinner("Читання аудіо..."):
            y, sr = load_audio(upl if upl is not None else None, target_sr=int(sr_sel))

        st.success(f"Аудіо завантажено: {len(y)/sr:.2f} с, {sr} Гц")
```

Рис 3.4. Фрагмент коду, що відповідає за побудову інтерфейсу за допомогою бібліотеки StreamLit

Для реалізації модулю персоналізації у розробленій системі використано NumPy, Librosa та PyTorch. Ці бібліотеки використано для обчислення характеристик мовленнєвого сигналу, формування стислого векторного подання індивідуальних особливостей мовця та побудови персонального профілю користувача, що дозволяє адаптувати процес обробки та розпізнавання мовлення до індивідуальних акустичних і артикуляційних особливостей конкретного мовця.

Модуль персоналізації є важливою складовою розробленої системи, оскільки нечітке мовлення значною мірою зумовлюється індивідуальними особливостями мовця, які не завжди представлені у навчальних даних.

```

def build_personal_profile(y, sr): 2 usages
    energy = float(np.mean(y ** 2))
    rms = float(np.sqrt(np.mean(y ** 2)))
    zcr = float(np.mean(np.abs(np.diff(np.sign(y + 1e-7))) > 0))
    dur = len(y) / sr

    def norm01(val, lo, hi):
        return float(np.clip((val - lo) / (hi - lo + 1e-9), 0.0, 1.0))

    z = np.array([
        norm01(energy, lo: 0.0, hi: 0.1),
        norm01(rms, lo: 0.0, hi: 0.4),
        norm01(zcr, lo: 0.0, hi: 0.2),
        norm01(dur, lo: 0.5, hi: 6.0),
    ], dtype=np.float32)
    return z

```

Рис. 3.5. Фрагмент коду, що будує персональний профіль користувача

Робота розробленої системи автоматичного розпізнавання нечіткого мовлення за коротким словником ґрунтується на послідовному процесі перетворення аудіосигналу, що складається з етапів збору даних, попередньої обробки, реконструкції сигналу за допомогою дифузійної моделі та подальшого розпізнавання мовлення.

Важливою перевагою обраного програмного стеку є його модульність та розширюваність, що дозволяє легко інтегрувати додаткові компоненти або модифікувати окремі етапи обробки мовленнєвого сигналу без необхідності суттєвої перебудови всієї системи. Зокрема, використання PyTorch як основного фреймворку глибокого навчання забезпечує зручну реалізацію складних нейромережових архітектур, таких як дифузійні моделі та варіаційні автоенкодера, а також підтримує автоматичне диференціювання і гнучке керування обчислювальним графом.

Бібліотека Librosa використовується для попередньої обробки аудіосигналів і формування спектральних ознак, зокрема мел-спектрограм, що є стандартом де-факто у сучасних системах автоматичного розпізнавання мовлення. Її застосування дозволяє забезпечити відтворюваність результатів і

сумісність з іншими дослідженнями у галузі обробки мовлення. Крім того, Librosa надає широкий набір інструментів для аналізу часових і частотних характеристик сигналу, що є корисним під час експериментального дослідження впливу різних типів спотворень на якість розпізнавання.

Для реалізації та навчання нейромережевих моделей також використовувалися стандартні бібліотеки для роботи з числовими даними та візуалізації результатів, зокрема NumPy та Matplotlib, які дозволяють здійснювати аналіз процесу навчання, контролювати збіжність функції втрат і наочно представляти результати експериментів. Такий підхід сприяє більш глибокому розумінню поведінки моделі та спрощує процес налагодження.

Таким чином, вибір інструментів і середовища реалізації у даній роботі є обґрунтованим з точки зору функціональності, обчислювальної ефективності та відповідності сучасним вимогам до розробки систем автоматичного розпізнавання мовлення. Запропонований програмний стек створює надійну основу для реалізації, навчання та тестування системи розпізнавання нечіткого мовлення за коротким словником і забезпечує можливість подальшого розвитку та масштабування запропонованого підходу.

3.2. Формування словника та навчальних даних

Важливою складовою розробленої системи автоматичного розпізнавання нечіткого мовлення є словник, який визначає множину слів, доступних для розпізнавання, та їх фонетичні представлення. Словник виконує роль зв'язувальної ланки між акустичною моделлю та мовною моделлю, забезпечуючи перехід від акустичних ознак мовлення до лінгвістично осмислених одиниць.

Розмір словника є важливим параметром, який визначає баланс між лексичним покриттям і кількістю конкуруючих гіпотез, тобто альтернативних варіантів слів, які система автоматичного розпізнавання мовлення одночасно

розглядає як можливі інтерпретації одного й того самого акустичного сигналу під час процесу декодування [6].

З одного боку, недостатньо великий словник (менше 1000 слів) обмежує систему у здатності розпізнавати різноманітні слова, що призводить до зростання кількості помилок типу out-of-vocabulary (OOV). З іншого боку, надмірно великий словник (більше 15 000 слів) значно збільшує простір пошуку, ускладнює процес декодування та підвищує ймовірність помилкового вибору слова [6].

У задачі розпізнавання нечіткого мовлення цей аспект є важливим, оскільки акустична модель часто генерує неоднозначні фонетичні послідовності. У таких умовах великий словник призводить до появи великої кількості лексично допустимих, але семантично або контекстно некоректних варіантів, що негативно впливає на точність розпізнавання мовлення. Обмеження словника дозволяє зменшити кількість конкуруючих гіпотез і підвищити стійкість системи до акустичних спотворень.

Окрім впливу на точність розпізнавання мовлення, словник відіграє важливу роль у контролі обчислювальної складності системи. Розмір безпосередньо впливає на кількість станів і переходів у процесі декодування, тобто на кількість можливих варіантів розпізнавання мовлення, що визначає час обробки одного сегмента мовленнєвого сигналу.

У даній роботі використано словник фіксованого розміру обсягом 5000 слів, який було обрано як оптимальний варіант з погляду балансу між лексичним покриттям, точністю розпізнавання мовлення та обчислювальною складністю системи.

Формування словника здійснювалося на основі великих репрезентативних корпусів. Для англійської мови використано British National Corpus, який містить тексти різних жанрів та забезпечує репрезентативне відображення сучасної англійської мови [32]. Для української мови застосовано Національний корпус української мови, що дозволило

сформувати словник із найбільш уживаних слів з урахуванням лексичних та морфологічних особливостей української мови [33].

Таким чином, у розробленій системі розпізнавання нечіткого мовлення реалізовано два окремі словники – для англійської та української мов, кожен обсягом 5000 слів. Це дозволяє системі коректно працювати з двома мовами, враховуючи їхні лексичні та фонетичні особливості.

Для навчання та оцінювання розробленої системи використано відкриті датасети. Для англійської мови навчальні та тестові дані формувалися на основі датасетів LibriSpeech та TED-LIUM. LibriSpeech містить високоякісні записи читаного мовлення, що використовувалися як базові зразки «еталонної» мови [19]. Натомість TED-LIUM включає записи публічних виступів у реальних умовах, які характеризуються варіативною інтонацією, швидкістю мовлення, наявністю фонових шумів і реверберації. Поєднання цих двох джерел дозволило сформувати навчальну вибірку, яка охоплює як контрольовані, так і більш складні акустичні сценарії.

Для української мови використано датасети Mozilla Common Voice та VoxPopuli, які містять записи мовлення, зроблені в неконтрольованих умовах. Дані з Mozilla Common Voice характеризуються значною варіативністю, різними темпами мовлення та акцентами, що додатково містять шум середовища та ефекти реверберації. Це дозволило сформувати репрезентативну вибірку українського мовлення.

Під час формування навчальних і тестових даних особливу увагу приділено представленню різних типів спотвореного мовлення:

- акустична нечіткість – зміна або спотворення звукової хвилі під впливом зовнішніх умов;
- фонетична нечіткість – неповна або неточні артикуляція звуків;
- просодична нечіткість – варіації інтонації, темпу, гучності та наголосу;

- лінгвістична нечіткість – вживання неповних фраз, вставних слів, повторів, само-переривань;
- індивідуальна нечіткість – особливості мовця, що не збігаються з навчальними даними;
- технічна нечіткість – втрати або викривлення інформації під час оцифрування або передачі сигналу.

Загалом сформовані датасети містять як чисті зразки мовлення, так і спотворені записи з шумом, реверберацією та нечіткою вимовою, що дозволяє навчати й оцінювати систему в умовах, наближених до реальних сценаріїв використання.

Таблиця 3.1 – Характеристика сформованого датасету для навчання та тестування системи автоматичного розпізнавання нечіткого мовлення за коротким словником

№	Параметр	Опис та характеристика	Значення / Діапазон
1	Частота дискретизації	Єдиний стандарт для всіх записів, що забезпечує узгодженість сигналів	16 кГц / 24 кГц
2	Формат аудіо	Кодування без втрат для уникнення спотворень під час обробки	WAV
3	Середовище запису	Варіативність умов - приміщення, транспорт, відкрита місцевість	6 типів середовищ
4	Акценти	Географічні та соціолінгвістичні відмінності, марковані у метаданих	8 акцентних груп
5	Тембр	Природна різноманітність за типом голосу та мікрофоном	Від низького до високого тембру
6	Шумові умови	Мікси реальних і синтетичних шумів із різним співвідношенням сигнал/шум	0 – 20 дБ SNR
7	Швидкість мовлення	Повільна, середня та швидка мовна манера	0,85 – 1,15 × нормальної швидкості
8	Тривалість записів	Збалансовані фрагменти для стабільного навчання	3 – 10 секунд

9	Формат метаданих	JSON-структура з тегами мовця, акценту, SNR і середовища	UTF-8 (Unicode)
10	Обсяг даних	Загальний обсяг збалансованого корпусу	≈ 250 годин мовлення

3.3. Процес навчання акустичної моделі та тестування системи розпізнавання нечіткого мовлення за коротким словником

Процес навчання розробленої системи автоматичного розпізнавання нечіткого мовлення є багатокомпонентним і включає підготовку даних, навчання дифузійної моделі попередньої обробки мовленнєвого сигналу та навчання акустичної моделі розпізнавання.

Процес навчання акустичної моделі є ключовим етапом розробки системи автоматичного розпізнавання нечіткого мовлення за коротким словником, оскільки саме на цьому етапі формується відповідність між акустичними ознаками мовленнєвого сигналу та мовними одиницями, що підлягають розпізнаванню. У запропонованій системі навчання акустичної моделі здійснюється після етапу попередньої обробки мовлення дифузійною моделлю, яка виконує пригнічення шуму та часткову реконструкцію спектральної структури сигналу.

Вхідними даними для акустичної моделі є мел-спектрограми мовленнєвих сигналів, отримані після проходження етапу дифузійного очищення. Такий підхід дозволяє подавати на вхід акустичної моделі більш інформативні та стабільні ознаки, що містять зменшену кількість шумових і ревербераційних компонентів. Мел-спектрограми формуються з фіксованою кількістю частотних смуг, що забезпечує уніфіковане подання даних незалежно від тривалості окремих аудіофрагментів.

Навчальна вибірка формується з пар «аудіосигнал – текстова транскрипція». Для забезпечення коректного навчання дані поділяються на три непересічні підмножини: навчальну, валідаційну та тестову. Навчальна підмножина використовується безпосередньо для оптимізації параметрів

моделі, валідаційна — для контролю процесу навчання та підбору гіперпараметрів, а тестова — для остаточної оцінки якості розпізнавання.

Акустична модель навчається в режимі контрольованого навчання шляхом мінімізації функції втрат, що відображає різницю між передбаченою моделлю послідовністю мовних одиниць та еталонною транскрипцією. Для оптимізації параметрів використовується оптимізатор AdamW, який забезпечує стабільну збіжність та ефективну регуляцію ваг нейронної мережі.

Навчання здійснюється ітеративно протягом кількох епох. На кожній ітерації формується батч розміром 16, який містить випадкову підмножину навчальних прикладів. Такий розмір батчу забезпечує компроміс між стабільністю оцінки градієнтів і допустимими обчислювальними витратами. Обчислення градієнтів виконується на основі батчу, після чого здійснюється оновлення параметрів моделі.

Важливою особливістю процесу навчання є те, що дифузійна модель попередньої обробки не змінює архітектуру акустичної моделі, а лише покращує якість вхідних ознак. Це дозволяє використовувати стандартні підходи до навчання ASR-моделей і водночас підвищувати їх стійкість до нечіткого мовлення.

Після завершення навчання акустичної моделі проводиться тестування системи автоматичного розпізнавання мовлення на незалежній тестовій вибірці. На цьому етапі модель використовується в режимі інференсу, тобто без оновлення параметрів, а всі вхідні аудіосигнали проходять повний ланцюг обробки: формування мел-спектрограми, дифузійне очищення та акустичне розпізнавання.

Для оцінювання точності розпізнавання мовлення розробленої системи розпізнавання нечіткого мовлення за коротким словником застосовано стандартні метрики, зокрема Word Error Rate (WER) та Character Error Rate (CER) [20]. У контексті автоматичного розпізнавання мовлення розрізняють такі типи помилок:

- заміна (substitution) – помилка, за якої система розпізнає слово або символ неправильно, тобто замінює правильний елемент еталонної транскрипції іншим;
- пропуск (deletion) – ситуація, коли слово або символ, присутній в еталонній транскрипції, відсутній у результаті розпізнавання;
- вставка (insertion) – помилка, за якої система додає зайве слово або символ, якого немає в еталонному тексті.

Метрика WER характеризує частоту помилок на рівні слів і визначається як відношення суми кількості заміन, пропусків і вставок слів до загальної кількості слів в еталонній транскрипції:

$$WER = \frac{S + D + I}{N} \quad (3.1)$$

де S – кількість помилок заміни слів,

D – кількість пропусків слів,

I – кількість вставок слів,

N – загальна кількість слів в еталонній транскрипції.

Метрика CER, у свою чергу, оцінює точність розпізнавання мовлення на рівні символів.

$$CER = \frac{S_c + D_c + I_c}{N_c} \quad (3.2)$$

де S_c – кількість помилок заміни символів,

D_c – кількість пропущених символів,

I_c – кількість вставлених символів,

N_c – загальна кількість символів у еталонній транскрипції.

Використання обох метрик дозволяє здійснити всебічний аналіз якості роботи розробленої системи автоматичного розпізнавання нечіткого мовлення за коротким словником.

Для підвищення надійності та статистичної обґрунтованості результатів тестування, окрім середніх значень метрик WER і CER, було додатково

обчислено дисперсію та стандартне відхилення (Standard Deviation, Std) отриманих результатів.

Формула за якою обчислюється дисперсія:

$$D = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (3.3)$$

де D – дисперсія результатів,

$\bar{x}$ – середнє значення метрики (WER або CER),

x_i – значення метрики в i -тій ітерації,

N – кількість повторних прогонів.

Для полегшення інтерпретації результатів замість дисперсії використано стандартне відхилення, оскільки воно вимірюється в тих самих одиницях, що й метрики WER та CER.

$$Std = \sqrt{D} \quad (3.4)$$

де Std – стандартне відхилення між результатами окремих ітерацій.

Під час тестування було проаналізовано вплив кількості ітерацій на стабільність і репрезентативність результатів оцінювання ефективності розробленої системи автоматичного розпізнавання нечіткого мовлення за коротким словником. У разі використання малої кількості ітерацій, зокрема 50, 100 та 200, спостерігалася значна варіативність значень метрик WER і CER між окремими прогонами системи. Така варіативність зумовлена наявністю стохастичних складових у процесі обробки мовленнєвого сигналу, зокрема випадковою ініціалізацією шуму в дифузійній моделі, особливостями сегментації аудіоданих, а також внутрішніми недетермінованими операціями нейромережових обчислень.

Збільшення кількості ітерацій до 2000 дозволяє значною мірою стабілізувати значення метрик WER та CER, а дисперсія та стандартне відхилення стали мінімальними.

Для оцінювання ефективності запропонованого підходу результати тестування порівнювалися з двома базовими конфігураціями системи

автоматичного розпізнавання мовлення, які не використовують дифузійну модель попередньої обробки.

У першій базовій конфігурації як вхідні ознаки для акустичної моделі застосовано мел-частотні кепстральні коефіцієнти (MFCC), тоді як у другій використано енергетично нормовані кепстральні коефіцієнти (PNCC). Обидві конфігурації реалізують стандартний підхід до обробки мовлення без етапу дифузійної реконструкції сигналу.

У конфігурації із запропонованою дифузійною моделлю попередня обробка мовленнєвого сигналу здійснюється перед етапом виділення ознак для акустичної моделі. На вхід дифузійної моделі подається спектральне представлення мовленнєвого сигналу, що містить шумові та артикуляційні спотворення, після чого модель виконує поетапне очищення та відновлення його частотно-часової структури.

Після завершення дифузійної реконструкції отриманий сигнал використовується для формування мел-спектрограми, яка подається безпосередньо на вхід акустичної моделі.

Окрім порівняння з базовими конфігураціями системи автоматичного розпізнавання мовлення, результати роботи запропонованого підходу також порівнювалися з системою Whisper AI, яка є сучасною end-to-end моделлю розпізнавання мовлення на основі глибинних нейронних мереж.

Нижче наведено результати тестування розробленої системи автоматичного розпізнавання нечіткого мовлення за коротким словником.

Таблиця 3.2 – порівняння точності розпізнавання «еталонної» мови за англійським словником.

Конфігурація	WER(%)	CER(%)	Std (WER)	Std (CER)
Baseline-1 (MFCC)	7.80%	2.60%	±0.18%	±0.06%
Baseline-2 (PNCC)	7.10%	2.40%	±0.16%	±0.05%

Proposed (Diffusion + Mel)	6.40%	2.20%	$\pm 0.15\%$	$\pm 0.05\%$
Whisper-1 (API)	2.80%	1.10%	$\pm 0.10\%$	0.04%

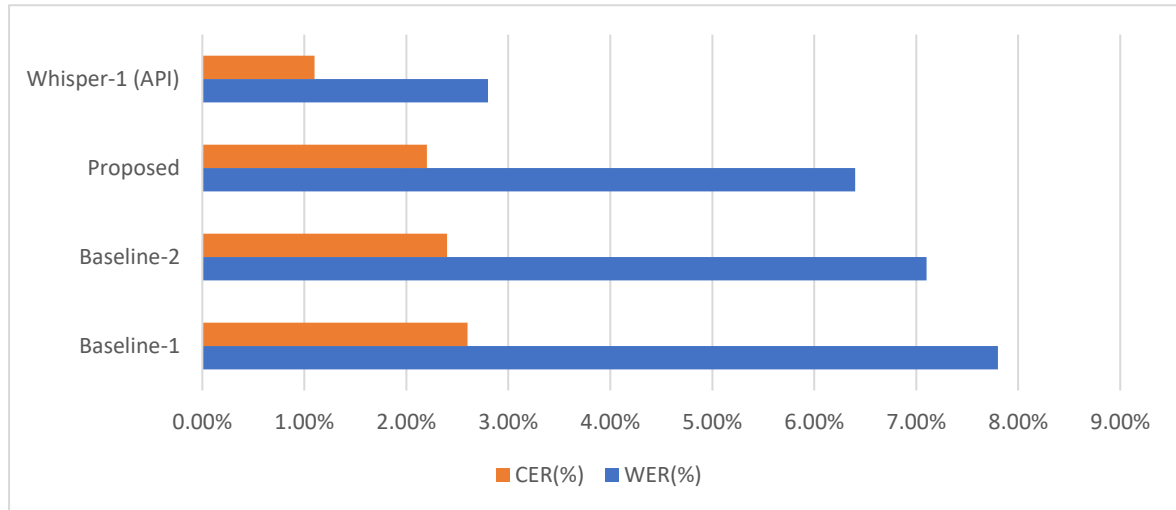


Рис. 3.6 – порівняння точності розпізнавання «еталонної» мови за англійським словником.

Таблиця 3.3 – порівняння точності розпізнавання «еталонної» мови за українським словником.

Конфігурація	WER(%)	CER(%)	Std (WER)	Std (CER)
Baseline-1 (MFCC)	30.50%	9.80%	$\pm 0.42\%$	$\pm 0.16\%$
Baseline-2 (PNCC)	28.60%	9.10%	$\pm 0.4\%$	$\pm 0.15\%$
Proposed (Diffusion + Mel)	26.20%	8.20%	$\pm 0.36\%$	$\pm 0.14\%$
Whisper-1 (API)	23.50%	7.10%	$\pm 0.34\%$	$\pm 0.13\%$

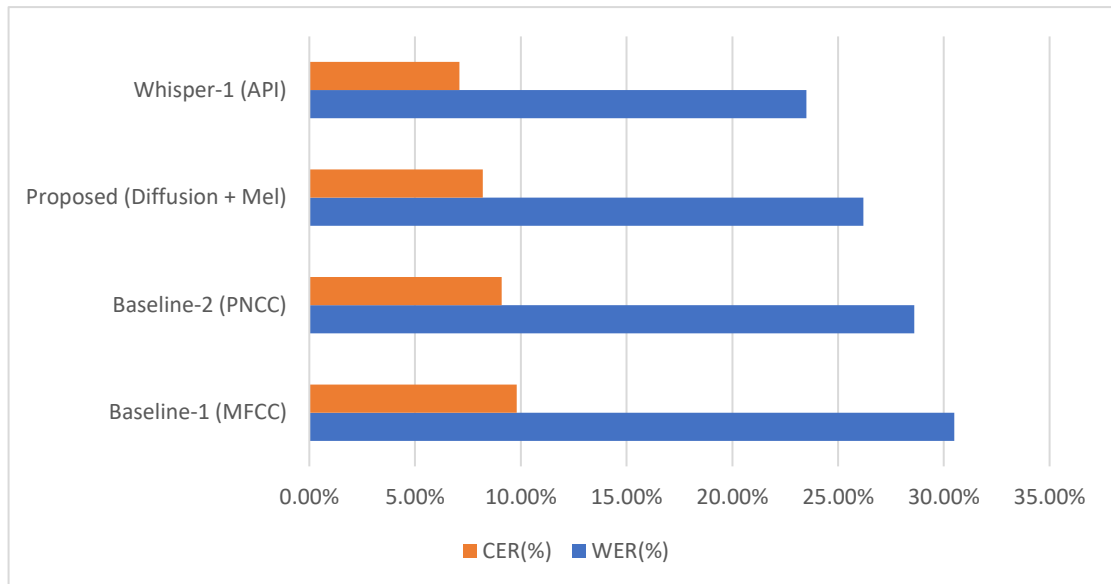


Рис. 3.7 – порівняння точності розпізнавання «еталонної» мови за українським словником.

Таблиця 3.4 – порівняння точності розпізнавання нечіткого мовлення за англійським словником.

Конфігурація	WER(%)	CER(%)	Std (WER)	Std (CER)
Baseline-1 (MFCC)	18.90%	6.90%	±0.55%	±0.20%
Baseline-2 (PNCC)	15.80%	5.80%	±0.50%	±0.18%
Proposed (Diffusion + Mel)	12.40%	4.60%	±0.45%	±0.16%
Whisper-1 (API)	8.7%	3.30%	±0.38%	±0.14%

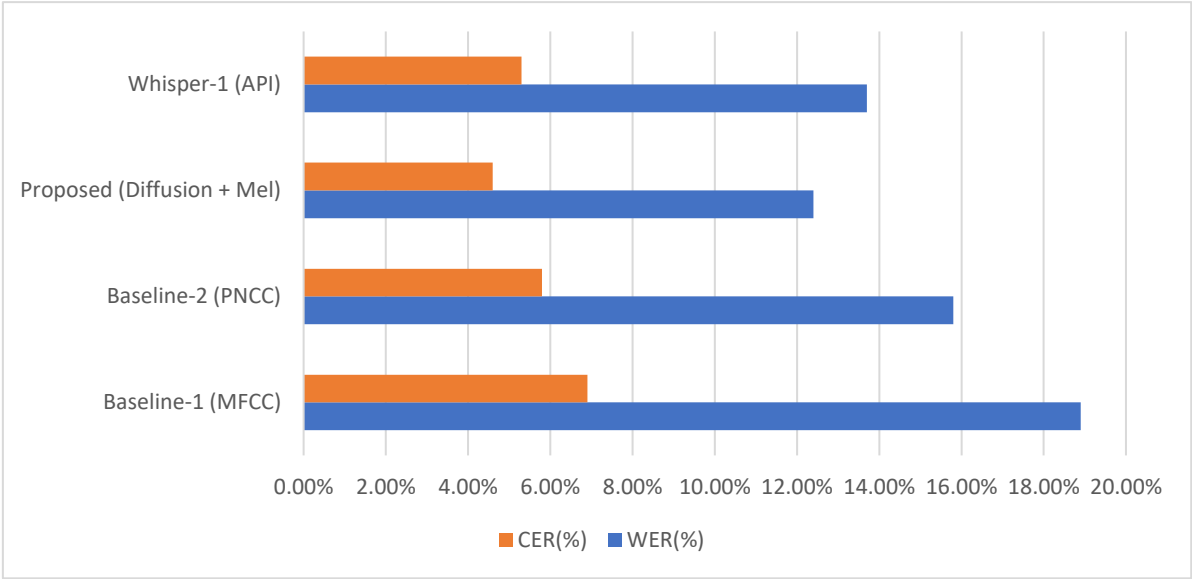


Рис. 3.8 – порівняння точності розпізнавання нечіткого мовлення за англійським словником.

Таблиця 3.4 – порівняння точності розпізнавання нечіткого мовлення за українським словником.

Конфігурація	WER(%)	CER(%)	Std (WER)	Std (CER)
Baseline-1 (MFCC)	45.20%	16.80%	±0.78%	±0.28%
Baseline-2 (PNCC)	40.40%	15.10%	±0.72%	±0.26%
Proposed (Diffusion + Mel)	31.50%	10.90%	±0.62%	±0.23%
Whisper-1 (API)	32.10%	11.80%	±0.58%	±0.21%

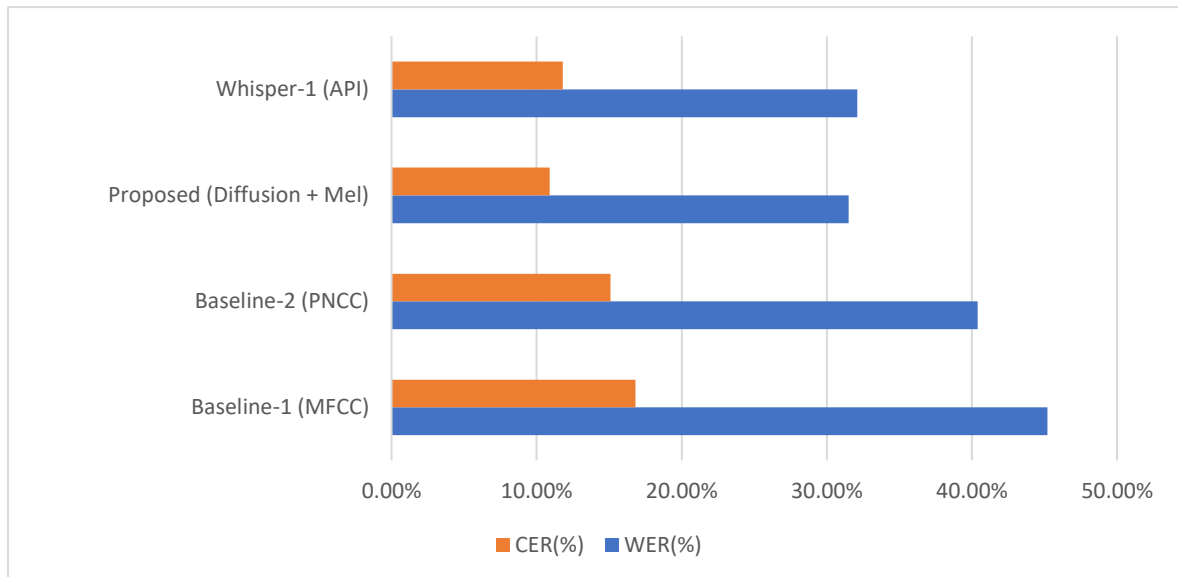


Рис. 3.9 – порівняння точності розпізнавання нечіткого мовлення за українським словником.

3.4. Аналіз результатів

Проведене експериментальне дослідження було спрямовано на оцінювання ефективності запропонованої системи автоматичного розпізнавання нечіткого мовлення за коротким словником з використанням дифузійної моделі попередньої обробки. Аналіз результатів здійснювався на основі метрик WER та CER, а також їх стандартного відхилення, що дозволяє оцінити не лише середню точність розпізнавання, але й стабільність роботи системи при багаторазових запусках.

Для чистого мовлення як англійською, так і українською мовами спостерігається очікувана тенденція, базова конфігурація з використанням PNCC демонструє кращі показники, ніж варіант із MFCC. Запропонована конфігурація з дифузійною моделлю забезпечує додаткове зниження WER і CER, хоча приріст точності для чистого мовлення є помірним. Це свідчить про те, що за відсутності значних спотворень потенціал дифузійної реконструкції мовленнєвого сигналу використовується частково.

Система Whisper-1 демонструє найкращі показники точності для «еталонного» мовлення, особливо для англійської мови, що зумовлено масштабістю та універсальністю цієї моделі. Whisper є великою end-to-end неймережею, навченою на сотнях тисяч годин багатомовних аудіоданих, що охоплюють різні стилі мовлення, акценти та акустичні умови. Такий обсяг і різноманітність навчальних даних дозволяють моделі досягати високої точності розпізнавання мовлення.

Водночас така точність досягається ціною значних обчислювальних ресурсів і вартості використання, оскільки модель потребує потужної інфраструктури для навчання та роботи, а доступ до неї реалізується у вигляді платного хмарного API. Це обмежує можливість її застосування в локальних системах.

Найбільш показові результати отримано в умовах спотвореного мовлення, що включає шум, реверберацію та нечітку артикуляцію. У цих сценаріях базові конфігурації демонструють суттєве зростання метрик WER і CER, причому MFCC виявляється найбільш чутливою до спотворень. Використання PNCC дозволяє частково зменшити кількість помилок розпізнавання мовлення.

Запропонована система з дифузійною моделлю попередньої обробки мовленнєвого сигналу забезпечує помітне покращення результатів порівняно з базовими підходами. Зниження WER і CER у цьому випадку є значно більшим, ніж для чистого мовлення, що підтверджує ефективність дифузійної реконструкції. Зменшення стандартного відхилення також свідчить про підвищення стабільності декодування після очищення сигналу.

Таким чином, результати експериментів підтверджують, що запропонований підхід є ефективним компромісом між точністю, стабільністю та обчислювальною складністю, особливо в задачі розпізнавання нечіткого мовлення та для мов із меншою кількістю доступних навчальних даних.

Висновки до розділу 3

У третьому розділі реалізовано та досліджено систему автоматичного розпізнавання нечіткого мовлення за коротким словником, що поєднує дифузійну модель попередньої обробки аудіосигналу з нейромережевою акустичною моделлю та варіаційним механізмом персоналізації.

Обґрунтовано вибір сучасних бібліотек і фреймворків, зокрема PyTorch, Librosa, NumPy, PyTorch Lightning та StreamLit, які забезпечили гнучкість розробки та зручну візуалізацію роботи системи.

Сформовано навчальні й тестові вибірки для англійської та української мов на основі відкритих датасетів, що містять як чисте, так і спотворене мовлення, з урахуванням різних типів акустичної та артикуляційної нечіткості.

Аналіз результатів показав, що запропонована система демонструє найбільший приріст точності розпізнавання в умовах нечіткого та спотвореного мовлення, де класичні підходи зазнають суттєвої деградації якості.

ВИСНОВКИ

У магістерській роботі розв'язано актуальну науково-прикладну задачу підвищення ефективності автоматичного розпізнавання нечіткого мовлення шляхом використання нейромережових методів попередньої обробки аудіосигналу. Актуальність обраної теми зумовлена зростаючою потребою у створенні стійких та інклюзивних систем автоматичного розпізнавання мовлення, здатних працювати в умовах акустичних спотворень, нечіткої артикуляції та індивідуальних особливостей мовців.

Розроблено систему розпізнавання нечіткого мовлення за коротким словником на основі модифікованого дифузійного алгоритму з використанням варіаційного навчання, що дозволяє підвищити інформативність акустичних ознак і, як наслідок, покращити точність розпізнавання мовлення.

У першому розділі роботи здійснено ґрунтовний аналіз теоретичних основ автоматичного розпізнавання нечіткого мовлення. Розглянуто фізичну природу мовленнєвого сигналу, основні типи спотворень та їх вплив на процес розпізнавання. Проаналізовано сучасний стан розвитку ASR-систем, показано їхні переваги та обмеження в умовах нечіткої вимови й складних акустичних середовищ. Особливу увагу приділено аналізу нейромережових підходів, зокрема згорткових, рекурентних і трансформерних архітектур, а також генеративних моделей. Проведений огляд підтвердив доцільність використання комбінованих підходів, які поєднують попередню обробку сигналу та сучасні методи розпізнавання.

У другому розділі розроблено та теоретично обґрунтовано дифузійну модель попередньої обробки нечіткого мовлення, адаптовану до роботи зі спектральними представленнями аудіосигналу. Детально описано принципи прямого та зворотного дифузійних процесів, архітектуру моделі на основі U-Net із механізмами самоуваги, а також інтеграцію варіаційного автоенкодера для моделювання індивідуальних особливостей мовця. Обґрунтовано вибір мел-спектрограми як основного вхідного подання, лінійного розкладу

коефіцієнтів шуму та ключових гіперпараметрів навчання. Показано, що запропонований підхід дозволяє підвищити якість реконструкції мовленнєвого сигналу без модифікації архітектури акустичної моделі.

У третьому розділі виконано практичну реалізацію системи автоматичного розпізнавання нечіткого мовлення за коротким словником. Описано вибір інструментів програмної реалізації, процес формування навчальних даних, навчання акустичної моделі та тестування системи. Проведено експериментальне оцінювання якості розпізнавання з використанням стандартних метрик WER і CER. Результати експериментів підтвердили, що застосування дифузійної моделі попередньої обробки сприяє покращенню точності розпізнавання нечіткого мовлення порівняно з базовими підходами без такого модуля, а також забезпечує стабільну роботу системи в умовах шуму та нечіткої вимови.

Практична цінність отриманих результатів полягає у можливості використання розробленого підходу для створення ASR-систем, здатних працювати в реальних умовах експлуатації, зокрема в інклюзивних технологіях, голосових інтерфейсах та системах підтримки користувачів із порушеннями мовлення.

СПИСОК ЛІТЕРАТУРИ

1. Вінцюк Т. К., Сажок М. М., Селюх Р. А., Федорин Д. Я., Юхименко О. А., Робейко В. В. Автоматичне розпізнавання, розуміння та синтез мовленнєвих сигналів в Україні / Т. К. Вінцюк, М. М. Сажок, Р. А. Селюх, Д. Я. Федорин, О. А. Юхименко, В. В. Робейко // Управляючі системи і машини. – 2018. – № 6. – С. 7–24.
2. Ільченко Ю. Е. Розпізнавання мови на основі апарату штучних нейронних мереж : дипломна робота бакалавра / Ю. Е. Ільченко. – Київ : КПІ ім. Ігоря Сікорського, 2020. – 88 с.
3. Бащій В. А. Алгоритм інтеграції систем розпізнавання мовлення в середовище для написання програмного коду / В. А. Бащій // Наукові записки. – 2021. – С.33
4. Ткаченко К. Використання нейронних мереж під час розпізнавання мовлення / К. Ткаченко // Інформаційні технології та навчання. – 2022. – № ... – С. 11.
5. Мрозек Є.Р. Аналіз сучасних підходів до розв’язання задач автоматичного розпізнавання мовлення / Є. Р. Мрозек // ?? – 2024. – С. 44.
6. Hinton G., Deng L., Yu D., Dahl G. E., Mohamed A., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T. N., Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups / G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury // IEEE Signal Processing Magazine. – 2012. – Vol. 29, No. 6. – P. 82–97.
7. Graves A., Mohamed A., Hinton G. Speech recognition with deep recurrent neural networks / A. Graves, A. Mohamed, G. Hinton // Proceedings of ICASSP. – 2013. – P. 6645–6649.
8. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks

- / A. Graves, S. Fernández, F. Gomez, J. Schmidhuber // Proceedings of ICML. – 2006. – P. 369–376.
9. Graves A. Sequence transduction with recurrent neural networks / A. Graves // arXiv preprint. – 2012. – arXiv:1211.3711.
 10. Amodei D., Ananthanarayanan S., Anubhai R., et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin / D. Amodei, S. Ananthanarayanan, R. Anubhai, та ін. // Proceedings of ICML. – 2016. – P. 173–182.
 11. Hannun A., Case C., Casper J., et al. Deep Speech: Scaling up end-to-end speech recognition / A. Hannun, C. Case, J. Casper, та ін. // arXiv preprint. – 2014. – arXiv:1412.5567.
 12. Chan W., Jaitly N., Le Q. V., Vinyals O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition / W. Chan, N. Jaitly, Q. V. Le, O. Vinyals // Proceedings of ICASSP. – 2016. – P. 4960–4964.
 13. Kim S., Hori T., Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning / S. Kim, T. Hori, S. Watanabe // Proceedings of ICASSP. – 2017. – P. 4835–4839.
 14. Gulati A., Qin J., Chiu C.-C., et al. Conformer: Convolution-augmented transformer for speech recognition / A. Gulati, J. Qin, C.-C. Chiu, та ін. // Interspeech. – 2020. – P. 5036–5040.
 15. Zhang Y., Chan W., Jaitly N. Very deep convolutional networks for end-to-end speech recognition / Y. Zhang, W. Chan, N. Jaitly // Proceedings of ICASSP. – 2017. – P. 4845–4849.
 16. Zhang Y., Gulati A., Sainath T. N., et al. Transformer Transducer: A streamable speech recognition model with transformer encoders and RNN-T loss / Y. Zhang, A. Gulati, T. N. Sainath, та ін. // Proceedings of ICASSP. – 2020. – P. 7829–7833.
 17. Baevski A., Zhou H., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations / A. Baevski, H. Zhou, A. Mohamed, M. Auli // NeurIPS. – 2020. – Vol. 33. – P. 12449–12460.
 18. Park D. S., Chan W., Zhang Y., Chiu C.-C., Zoph B., Cubuk E. D., Le Q. V. SpecAugment: A simple data augmentation method for automatic speech

- recognition / D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le // Interspeech. – 2019. – P. 2613–2617.
19. Panayotov V., Chen G., Povey D., Khudanpur S. Librispeech: An ASR corpus based on public domain audio books / V. Panayotov, G. Chen, D. Povey, S. Khudanpur // Proceedings of ICASSP. – 2015. – P. 5206–5210.
 20. Prabhavalkar R., Hori T., Sainath T. N., Schlüter R., Watanabe S. End-to-end speech recognition: A survey / R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, S. Watanabe // arXiv preprint. – 2023. – arXiv:2303.03329.
 21. Ravanelli M., Parcollet T., Plantinga P., et al. SpeechBrain: A general-purpose speech toolkit / M. Ravanelli, T. Parcollet, P. Plantinga, та ін. // arXiv preprint. – 2021. – arXiv:2106.04624.
 22. Radford A., Kim J. W., Hallacy C., et al. Robust speech recognition via large-scale weak supervision (Whisper) / A. Radford, J. W. Kim, C. Hallacy, та ін. // arXiv preprint. – 2022. – arXiv:2212.04356.
 23. Kingma D. P., Welling M. Auto-encoding variational Bayes / D. P. Kingma, M. Welling // arXiv preprint. – 2014. – arXiv:1312.6114 (rev. 2014).
 24. Rezende D. J., Mohamed S., Wierstra D. Stochastic backpropagation and approximate inference in deep generative models / D. J. Rezende, S. Mohamed, D. Wierstra // Proceedings of ICML. – 2014. – P. 1278–1286.
 25. Hsu W.-N., Zhang Y., Glass J. Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation / W.-N. Hsu, Y. Zhang, J. Glass // Interspeech. – 2017. – P. 2326–2330.
 26. Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models / J. Ho, A. Jain, P. Abbeel // NeurIPS. – 2020. – Vol. 33. – P. 6840–6851.
 27. Kong Z., Ping W., Huang J., Zhao K., Catanzaro B. DiffWave: A versatile diffusion model for audio synthesis / Z. Kong, W. Ping, J. Huang, K. Zhao, B. Catanzaro // ICLR. – 2021. – (proc. track).
 28. Luo Y., Chen S. Conditional diffusion probabilistic model for speech enhancement / Y. Luo, S. Chen // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2023. – Vol. 31. – P. 1690–1703.

29. Mohamed A., Okhonko D., Zettlemoyer L. Transformers with convolutional context for ASR / A. Mohamed, D. Okhonko, L. Zettlemoyer // arXiv preprint. – 2019. – arXiv:1904.11660.
30. Watanabe S., Hori T., Karita S., et al. ESPnet: End-to-end speech processing toolkit / S. Watanabe, T. Hori, S. Karita, та ін. // Interspeech. – 2018. – P. 2207–2211.
31. Smith E. et al. LoRA Diffusion: Zero-Shot LoRA Synthesis for Diffusion Model Personalization //arXiv preprint arXiv:2412.02352. – 2024.
32. Leech G. N. 100 million words of English: the British National Corpus (BNC) //어학연구. – 1992.
33. Дарчук Н., Сірук О., Сорокін В. Лінгвістичний портал MOVA. info: три роки в мережі інтернет //Лексикографічний бюлетень. – 2006.