Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя
(повне найменування вищого навчального закладу)
Факультет комп'ютерно-інформаційних систем і програмної інженерії
(назва факультету)
Кафедра комп'ютерних систем та мереж
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістра
(освітній ступінь))

на тему: **Методи та інструменти побудови комп'ютерних систем аналізу і перетворення текстових повідомлень в аудіопотік**

Виконав: студент (ка) 6 курсу, групи СІм-61
спеціальності 123 «Комп'ютерна інженерія»
(шифр і назва спеціальності)

	(підпис)	Макогон С.В. (прізвище та ініціали)
Керівник	(підпис)	Луцків А.М. (прізвище та ініціали)
Нормоконтроль	(підпис)	Луцик Н.С. (прізвище та ініціали)
Завідувач кафедри	(підпис)	Осухівська Г.М. (прізвище та ініціали)
Рецензент	(підпис)	Стадник М.А. (прізвище та ініціали)

Тернопіль
2023

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя
(повне найменування вищого навчального закладу)
Факультет комп'ютерно-інформаційних систем і програмної інженерії
Кафедра комп'ютерних систем та мереж

ЗАТВЕРДЖУЮ
Завідувач кафедри *Осухівська Г.М.*

«_____» 2023 р.

З А В Д А Н Н Я

НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня *магістр*
(назва освітнього ступеня)

за спеціальністю *123 «Комп'ютерна інженерія»*
(шифр і назва спеціальності)

студенту *Макогон Сергій Віталійович*
(прізвище, ім'я, по-батькові)

1. Тема проекту (роботи) *Методи та інструменти побудови комп'ютерних систем аналізу і перетворення текстових повідомлень в аудіопотік*

Керівник проекту (роботи) *Луцків Андрій Мирославович, к.т.н., доц.*
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від «*01*» *грудня* 2023 року №*4/7-1132*

2. Термін подання студентом завершеної роботи

3. Вихідні дані до роботи *Моделі перетворення тексту в аудіо, типи TTS систем, параметри ефективності TTS систем*

4. Зміст роботи (перелік питань, які потрібно розробити)
Вступ. 1. Аналіз підходів до синтезу голосових повідомлень 2. Моделі та алгоритми побудови акустичних моделей 3. Імплементація комп'ютерної системи перетворення тексту в аудіопотік на основі Raspberry Pi 4. Охорона праці та безпека в надзвичайних ситуаціях. Висновки

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
1.Актуальність і мета дослідження. 2. Задачі, об'єкт і предмет, наукова новизна і практична цінність дослідження. 3. Характеристики аудіосигналів. 4. Методи перетворення тексту в аудіопотік 5. Архітектура трансформерів. 6. Структура вокодера. 7. Структура системи на Raspberry PI. 8. Результати перетворення тексту в аудіопотік 8. Висновки

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці та	*Осухівська Г.М., зав. каф. КС*		
безпека в надзвичайних			
ситуаціях	*Стадник І.Я., проф. каф. ОХ*		

7. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	*Аналіз підходів до синтезу голосових повідомлень*	*01.12.2023-*	*виконано*
		05.12.2023	
2.	*Моделі та алгоритми побудови акустичних моделей*	*05.12.2023-*	*виконано*
		12.12.2023	
3.	*Імплементація комп'ютерної системи перетворення*	*12.12.2023-*	*виконано*
	тексту в аудіопотік на основі Raspberry Pi 4	*17.12.2023*	
4.	*Охорона праці та безпека в надзвичайних ситуаціях*	*18.12.2023*	*виконано*
5.	*Оформлення пояснювальної записки*	*20.12.2023*	*виконано*
6.	*Оформлення графічного матеріалу*	*21.12.2023*	*виконано*
7.	*Попередній захист кваліфікаційної роботи магістра*	*22.12.2023*	*виконано*
8.	*Захист кваліфікаційної роботи магістра*		

Студент _____ *Макогон С.В.*
(підпис) (прізвище та ініціали)

Керівник проекту (роботи) _____ *Луцків А.М.*
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Методи та інструменти побудови комп'ютерних систем аналізу і перетворення текстових повідомлень в аудіопотік // Кваліфікаційна робота магістра// Макогон Сергій Віталійович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп’ютерно-інформаційних систем та програмної інженерії, група СІм-61 // Тернопіль, 2023 // с. – 86 , рис. – 33 , табл. – 8 , аркушів А1 –8 , додат. – 1, бібліогр. – 23.

Ключові слова: метод, інструмент, комп’ютерна система, перетворення, текст, аудіопотік.

У кваліфікаційній роботі магістра на основі аналізу таксономії процесів перетворення текстових повідомлень в аудіопотік визначено потенційні способи розвитку існуючих нейромережевих моделей, зокрема, в контексті застосування методів машинного навчання для підвищення якості попереднього опрацювання тексту, перетворення графем у фонеми, а також забезпечення можливості їх прогнозування на основі попередньо навчених нейронних моделей.

Запропоновано архітектуру нейронної мережі до складу якої входить енкодер на базі трансформерів, які забезпечують зменшення розмірності вхідної матриці фонем у 4 рази та добувають фонетичні властивості, а також декодер, який сформований з блоків визначення акустичних властивостей, зокрема енергії, тривалості і висоти звуку, та блоків декодування властивостей аудіосигналу.

Розроблено системний програмний додаток для забезпечення трансляції текстових повідомлень в аудіосигнал з використанням мови програмування Python та проведено експерименти на Raspberry PI 4.

ABSTRACT

Methods and instruments for building computer systems for analyzing and transforming text messages into audio streams /Master thesis / Makohon Serhii / Ternopil Ivan Pul'uj National Technical University, Faculty of Computer Information Systems and software engineering, group CIm -61 // Ternopil, 2023// p. - 86, fig. – 33, table. – 8, Sheets A1 – 8, Add – 1, Ref. – 23.

Keywords: method, tool, computer system, transforming, text, audio stream.

In the master's qualification work, based on the analysis of the taxonomy of the processes of converting text messages into an audio stream, potential ways of developing existing neural network models are identified, in particular, in the context of the application of machine learning methods to improve the quality of pre-processing of text, converting graphemes into phonemes, as well as ensuring the possibility of their prediction based on pre-trained neural models.

The architecture of a neural network is proposed, which includes an encoder based on transformers that provide a 4-fold reduction in the dimensionality of the input phoneme matrix and extract phonetic properties, as well as a decoder that is formed from blocks for determining acoustic properties, in particular energy, duration and pitch of sound, and blocks decoding of audio signal properties.

A system software application was developed to ensure the translation of text messages into an audio signal using the Python programming language, and experiments were conducted on Raspberry PI 4.

ЗМІСТ

ВСТУП

Актуальність теми. Сучасна динаміка розвитку інформаційних технологій особливо у сферах опрацювання великих даних, інтелектуального аналізу інформації, застосування IoT пристроїв і хмарних рішень сприяє розв'язку задач чи навіть проблем, які до цього часу залишались не розв'язаними, або потребували значних фінансових ресурсів.

Симбіоз IoT технологій та штучного інтелекту дає змогу автоматизувати багато різноманітних процесів у сфері медицини, зокрема, підвищити якість життя незрячим шляхом застосування електронних голосових помічників.

У динамічному ландшафті цифрового спілкування технологія синтезу мовлення стала ключовим інструментом із далекосяжними наслідками. Ця трансформаційна технологія дозволяє перетворювати текст у голосові повідомлення, розблоковуючи доступність, ефективність та інновації в різних сферах.

Голосові повідомлення є одним із основних засобів спілкування. Якщо пристрої можуть синтезувати природну мову на основі тексту, то новий тип взаємодії з електронними гаджетами та приладами реально впровадити у теперішнє життя. Якщо пристрої зможуть виконувати автономний синтез голосових повідомлень, не покладаючись при цьому на хмарні служби, то це призведе до появи нових застосувань і переваг такої технології. Наприклад, маршрутизатор WiFi може повідомити користувача, про проблеми з доступом до мережі Інтернет. Розумна камера, встановлена у віддаленому районі, може попередити злочини. Ці корисні дії виконуються пристроєм автономно й без використання хмарних служб. Додатковими перевагами синтезу голосових повідомлень, які генеруються на пристроях IoT є пом'якшення проблем конфіденційності, підвищення надійності та гарантована висока швидкість реагування, низька затримка та доступність. Тому, дослідження методів і засобів синтезу, а також

перетворення текстових повідомлень у голосові з реалізацією комп'ютерної системи з використанням IoT пристроїв є доволі актуальною науково-практичною задачею.

Розробка комп'ютерних систем перетворення тексту у звук вимагає знань та аналізу існуючих досліджень про особливості формування людського мовлення, а також включає в себе кілька напрямів, зокрема, лінгвістику, акустику, цифрову обробку сигналів і машинне навчання.

Комплексному дослідженню проблем перетворення текстових повідомлень у голосові з використанням сучасних технологій присвячено багато праць як вітчизняних, так і закордонних вчених. Серед вітчизняних науковців потрібно виділити результати, одержані вченими з інституту кібернетики НАН України (І. Крак, С. Кондратюк, В. Резник, Б. Гінзбург), та ряду інших – Р. Багрій, В. Ковтун, О. Бармак, І. Кривонос і т.д. Серед закордонних учених – Y. Ren, T. Qin, J. Shen, S.O. Arik, K. Kumar та ін.

Проте, зважаючи на серйозність та важливість отриманих наукових і практичних результатів, все ж залишається багато задач у сфері перетворення тексту у звук, зокрема щодо оптимізації алгоритмів та методів опрацювання тексту, синтезу звуку на IoT пристроях.

Мета кваліфікаційної роботи полягає у дослідженні та оптимізації існуючих методів, алгоритмів та засобів перетворення текстових повідомлень у голосові з використанням IoT і моделей машинного навчання.

Серед сукупності задач, які потрібно вирішити у кваліфікаційній роботі магістра, основними є:

– аналітичний огляд існуючих технологій перетворення тексту у звук;

– дослідження особливостей аналізу тексту і синтезу голосових повідомлень;

– побудова та оптимізація моделі трансформації тексту у звук на основі нейромережевого підходу;

– розробка методу, алгоритмів і процедур перетворення текстових повідомлень у голосові на базі Raspberry PI;

– налаштування апаратної і програмної складової комп'ютерної системи перетворення тексту у звук;

– розробка програмного забезпечення підтримки методу перетворення текстових повідомлень у голосові.

Об'єкт дослідження: процеси аналізу текстової інформації та синтезу голосових повідомлень.

Предмет дослідження: методи і засоби перетворення текстових даних у голосові повідомлення.

Методи дослідження: Для вирішення поставлених у кваліфікаційній роботі задач використано методи: аналізу та узагальнення – при дослідженні характеристик аудіосигналів та попередньому опрацюванні текстової інформації, нейромережевих технологій і машинного навчання – при побудові архітектури моделі перетворення тексту в аудіопотік; проектування, програмування – при розробці програмного забезпечення трансляції і відтворення текстових файлів у звуковий сигнал; експеримент та вимірювання – при визначенні ефективності запропонованого рішення.

Наукова новизна отриманих результатів. Наукова новизна, одержаних у роботі результатів полягає в наступному.

– уперше запропоновано архітектуру нейронної мережі до складу якої входить енкодер на базі трансформерів, які забезпечують зменшення розмірності вхідної матриці фонем у 4 рази та добувають фонетичні властивості, а також декодера, який сформований з блоків визначення акустичних властивостей, зокрема енергії, тривалості і висоти звуку, та блоків декодування властивостей аудіосигналу на основі якого відбувається генерація спектрограми Мела, що дало змогу забезпечити високу продуктивність і якість формування голосового повідомлення у порівнянні з іншими моделями.

– уперше розроблено системний програмний додаток для забезпечення трансляції текстових повідомлень в аудіосигнал з використанням мови програмування Python, що дало змогу відправляти текстові файли з користувацького пристрою та опрацьовувати їх вміст за допомогою запропонованої нейронної мережі і відтворювати результат у вигляді аудіосигналу на Raspberry PI.

Практичне значення одержаних результатів. Програмно реалізовано модель запропонованої архітектури нейронної мережі та проведено експериментальне порівняння з іншими моделями, що дало можливість довести доцільність реалізації системи, оскільки у запропонованій моделі на порядок менше параметрів та існує можливість ефективно функціонувати на пристроях з обмеженими ресурсами, зокрема Raspberry PI, починаючи з версії 2.

Публікації. Результати кваліфікаційної роботи апробовані на XII Міжнародній науково-практичній конференції молодих учених та студентів (6-7 грудня 2023 р.) та XI науково-технічній конференції Тернопільського національного технічного університету імені Івана Пулюя «Інформаційні моделі, системи та технології» (13-14 грудня 2023 року) як тези конференцій.

1. Луцків А.М., Макогон С.В. Нейромережеві підходи до перетворення текстових повідомлень в аудіопотік. Матеріали XII міжнародної науково-практичної конференції молодих учених та студентів «Актуальні задачі сучасних технологій» (6-7 грудня 2023 року). Тернопіль: ТНТУ. 2022. С. 438.

2. Луцків А.М., Макогон С.В. Типи архітектур нейронних мереж для перетворення текстових повідомлень у звуковий потік. Матеріали XI науково-технічної конференції Тернопільського національного технічного університету імені Івана Пулюя «Інформаційні моделі, системи та технології» (13-14 грудня 2023 року). Тернопіль: ТНТУ. 2022. С. 164.

Структура роботи. Кваліфікаційна робота містить розрахунково-пояснювальну записку та графічний матеріал. До складу записки входить вступу, 4 розділи, загальні висновки, список використаних джерел і додатки. Обсяг роботи: розрахунково-пояснювальна записка – 86 арк. формату А4, графічна частина – 8 аркушів формату А1.

РОЗДІЛ 1
АНАЛІЗ ПІДХОДІВ ДО СИНТЕЗУ ГОЛОСОВИХ ПОВІДОМЛЕНЬ

1.1. Аналіз підходів синтезу голосових повідомлень

Люди намагалися побудувати машини для синтезу людської мови ще в 12 столітті [1]. У другій половині 18 століття угорський вчений Вольфганг фон Кемпелен сконструював машину, яка забезпечувала генерацію звуків та містила множину міхів, пружин, волинки та резонансні коробки для створення простих слів і коротких речень [2].

Перша система синтезу голосових повідомлень, яка реалізована за допомогою комп'ютерів, з'явилася в другій половині 20 століття [3]. Ранні методи комп'ютерного синтезу мовлення включають артикуляційний синтез [4-6], формантний синтез [7-9] і конкатенативний синтез [10-13]. Пізніше, як розвиток статистичного машинного навчання, пропонується статистичний параметричний синтез мовлення (SPSS) [14-16], який передбачає аналіз таких параметрів, як спектр, несуча частота та тривалість для синтезу мовлення.

З 2010-х років синтез мовлення на основі нейронних мереж поступово став домінуючим методом і досяг значно кращої якості при генерації голосових повідомлень.

Артикуляційний синтез генерує мовлення, імітуючи поведінку людського артикулятора, такого як губи, язик, голосова щілина та рухомий голосовий тракт. В ідеалі артикуляційний синтез може бути найефективнішим методом синтезу мовлення, оскільки це спосіб, яким людина створює мову. Однак на практиці дуже важко змоделювати таку поведінку артикулятора. Наприклад, важко зібрати дані для його моделювання.

Тому якість мовлення при артикуляційному синтезі зазвичай гірша, ніж при пізнішому формантному синтезі та конкатенативному синтезі.

Синтез формант формує мовлення на основі набору правил, які керують спрощеною моделлю джерело-фільтр. Ці правила, зазвичай, розробляються лінгвістами, щоб якомога точніше імітувати структуру формантів та інші спектральні властивості мови. Голос синтезується модулем додаткового синтезу та акустичною моделлю зі змінними параметрами, такими як несуча частота, голос і рівні шуму.

Формантний синтез може забезпечити високий рівень розбірливості мовлення і при цьому використовується не надто значні обчислювальні ресурси, які добре підходять для вбудованих систем, і не покладаються на великомасштабний людський мовний корпус, як при конкатенативному синтезі. Проте синтезована мова звучить менш природно та має артефакти. Крім того, важко вказати правила синтезу.

Конкатенативний синтез ґрунтується на конкатенації фрагментів мови, які зберігаються в базі даних. Зазвичай, база даних складається з мовних одиниць, починаючи від цілого речення до складів, які записуються акторами голосу. У результаті застосовування цього підходу, конкатенативна система перетворення тексту у голосові повідомлення шукає мовні одиниці відповідно до заданого введеного тексту та створює форму мовного сигналу шляхом об’єднання цих одиниць разом. Загалом, конкатенативний підхід може створювати аудіо з високою розбірливістю та автентичним тембром, близьким до оригінального голосу актора. Однак, конкатенативний TTS вимагає величезної бази даних запису, щоб охопити всі можливі комбінації мовленнєві одиниці при вимові слів. Іншим недоліком є те, що створений голос є менш природним емоційний, оскільки конкатенація може призвести до меншої плавності наголосу, емоцій, просодії тощо.

Для усунення недоліків конкатенативного перетворення тексту у мовлення, пропонується статистичний параметричний синтез мови (англ. SPSS). Основна ідея полягає в тому, що замість прямого генерування хвилі за допомогою конкатенації можна спочатку згенерувати акустичні

параметри, необхідні для створення мовлення, а потім відновити його зі згенерованих акустичних параметрів за допомогою деяких алгоритмів [17-19].

SPSS зазвичай складається з трьох компонентів: модуль аналізу тексту, модуль прогнозування параметрів (акустична модель) і модуль аналізу/синтезу вокодера (вокодер).

Модуль аналізу тексту спочатку обробляє текст, включаючи нормалізацію тексту, перетворення графем у фонеми, сегментацію слів тощо, а потім витягує лінгвістичні характеристики, такі як фонеми, тривалість і POS-теги з різним рівнем деталізації.

Акустичні моделі (наприклад, на основі прихованої моделі Маркова (ПММ) навчаються з парними лінгвістичними ознаками та параметрами (акустичними ознаками), де акустичні характеристики включають основну частоту, спектр тощо, і витягуються з мовлення через вокодерний аналіз. Вокодери синтезують мову з прогнозованих акустичних характеристик.

SPSS має кілька переваг перед попередніми системами перетворення тексту у голосові повідомлення:

– природність аудіо;

– гнучкість та зручність налаштування параметрів для керування генерованим мовленням;

– низька вартість даних, тобто менша кількість записів, ніж при конкатенативному синтезі.

Однак SPSS також має свої недоліки:

– згенероване мовлення має нижчу розбірливість через артефакти, такі як глухий, дзижчачий або шумний звук;

– згенерований голос все ще є роботизованим і його можна легко відрізнити від людського запису мови.

У 2010-х роках, нейронні мережі та глибоке навчання досягли швидкого прогресу, що дало змогу застосувати їх для розв’язку задач, пов’язаних з генерацією голосових повідомлень з тексту. При цьому високу

ефективність показали нейронні мережі DNN [18] і рекурентні нейронні мережі RNN [19].

Однак ці моделі передбачають опрацювання акустичних характеристик на основі лінгвістичних особливостей, які відповідають парадигмі SPSS.

Пізніше Wang та ін. пропонують безпосередньо генерувати акустичні характеристики з послідовності фонем замість лінгвістичних ознак, що можна розглядати як перше дослідження для наскрізного синтезу мовлення.

Як розвиток парадигм при перетворенні тексту у голосові повідомлення на основі нейронних мереж, з'являються (глибокі) нейронні мережі як базові моделі для синтезу мовлення.

Деякі ранні нейронні моделі прийняті в статистичних параметричних моделях, які прийшли на заміну прихованих марковських моделей почали використовуватися при акустичному моделюванні. Пізніше було запропоновано WaveNet [20] для прямого генерування хвилі з лінгвістичних особливостей, що можна розглядати як першу сучасну нейронну модель TTS.

Інші моделі, такі як DeepVoice ver.1/ ver.2, як і раніше дотримуються трьох компонентів статистичного параметричного синтезу, але оновлюють їх відповідними моделями на основі нейронної мережі.

Крім того, пропонуються деякі наскрізні моделі (наприклад, Tacotron, Deep Voice 3 і FastSpeech) для спрощення модулів аналізу тексту та безпосереднього використання послідовності символів/фонем як вхідних даних та спрощення акустичних характеристик за допомогою мел-спектрограм.

Основні компоненти, які використовуються при застосуванні підходу нейронних мереж перетворення тексту у голосові повідомлення показано на рис. 1.1.

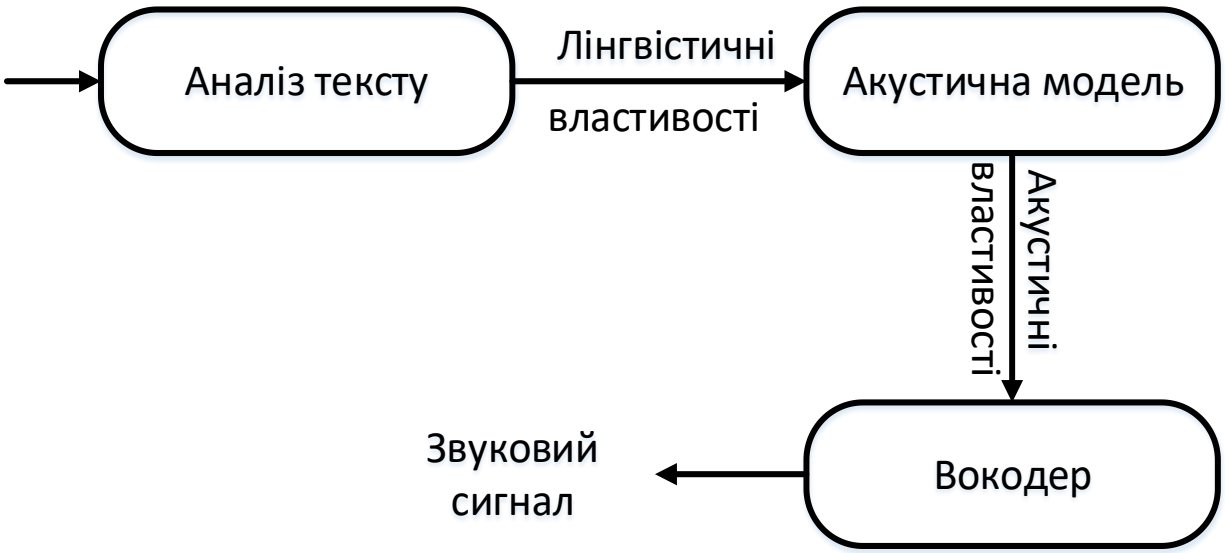

Рис. 1.1. Основні компоненти та алгоритм перетворення тексту у голосові повідомлення при використанні нейронних мереж

Пізніше були розроблені повністю наскрізні системи перетворення тексту в аудіо для безпосереднього генерування сигналу з тексту, такі як ClariNet, FastSpeech 2s і EATS. Порівняно з попередніми системами, заснованими на конкатенативному синтезі та статистичному параметричному синтезі, переваги синтезу мовлення на основі нейронної мережі включають високу якість голосу з точки зору як розбірливості, так і природності, а також менші вимоги до попередньої обробки людиною та розробки функцій.

Перетворення текстової інформації у голосові повідомлення має багато прикладних застосувань, зокрема, це стосується широко поширеної технології орієнтованої на допомогу людям із різними вадами зору. Окрім цього, перетворення тексту у звук можна застосовувати у сфері розваг з метою здешевлення озвучування фільмів, ігор та інших подібних речей.

Сьогодні існує декілька високоякісних рішень для перетворення тексту у звук [1,2], однак вони зазвичай забезпечують високу якість для дуже найбільш поширених мов, зокрема, англійської, французької, іспанської та ін.

Синтез звукових повідомлень для менш поширених мов, в тому числі української, є значно складнішим завданням, оскільки практично відсутні набори даних для проведення навчання засобами машинного навчання.

При цьому актуальними задачами є організація фреймворку з відкритим вихідним кодом, який містив би конвеєр для зручного навчання нових моделей синтезу звуку українськомовного контенту. При цьому критично важливо забезпечити ефективність процедури перетворення тексту у голосові повідомлення за допомогою тестових зразків голосу.

Для вирішення таких задач потрібно досліджувати семантичний та лінгвістичний аналіз [3], включно з фонетичним аналізом і граматичною структурою мови [4-6], розмічуванням і сегментацію мовних сигналів [7].

На сьогодні існує лише кілька рішень, які дають змогу формувати перетворення тексту українською мовою у голосові повідомлення. Більшість з них представляють собою моделі формантного синтезу в основі яких лежить підхід генерації на основі правил. Такі підходи мають перевагу низького використання ресурсів і високошвидкісного синтезу за рахунок синтезу штучного роботоподібного мовлення. Вони зазвичай використовуються як програми для читання з екрана для людей із вадами зору. Також існує два рішення для глибокого навчання української мови: сервіс на базі WaveNet від Google і Nuance Vocalizer. Однак вони є закритими, і забезпечити навчання нейронної мережі за допомогою нових наборів даних неможливо. Синтез мовлення з тексту за один наскрізний крок є дуже складним завданням, таку модель важко навчати та інтерпретувати.

Зазвичай, синтез голосових повідомлень поділяють на два етапи. Перший етап полягає в створенні синтезатора (який також відомий як архітектура кодера-декодера на послідовність символів, щоб передбачити мел-спектрограми [6] (акустичне представлення низького рівня).

Другий етап передбачає навчання моделі нейронного вокодера (генератора хвилі), який використовує акустичні характеристики з

попереднього кроку для реконструкції аудіосигналів (остаточний зразок голосу).

Щоб отримати зменшену розмірність, варто використовувати швидке перетворення Фур'є. Це стосується деяких даних, наприклад, аудіо чи інших сигналів. Використовуючи таке нелінійне перетворення, можна спостерігати базову природу даних, зосереджуючись переважно на низькочастотних і більш виразних деталях сигналу, а не на високочастотних деталях, які можуть містити деякі артефакти через процес дискретизації або розмір вікна.

При реалізації цих етапів пропонується використовувати сучасні моделі архітектури глибокого навчання – Tacotron2 для генерації мел-спектрограми та Parallel WaveGan для генерації аудіо.

1.2. Аналіз основної таксономії процесу перетворення тексту у голосові повідомлення з використанням нейронних мереж

Застосування нейромережевого підходу для перетворення тексту в аудіо можна класифікувати в основному з точки зору основних компонентів: аналіз тексту, акустичні моделі, вокодери та повністю наскрізні моделі, як показано на рис 1.2. Вони формують таксономію та основні поняття при побудові алгоритмів щодо практичної реалізації методів перетворення тексту у голосові повідомлення.

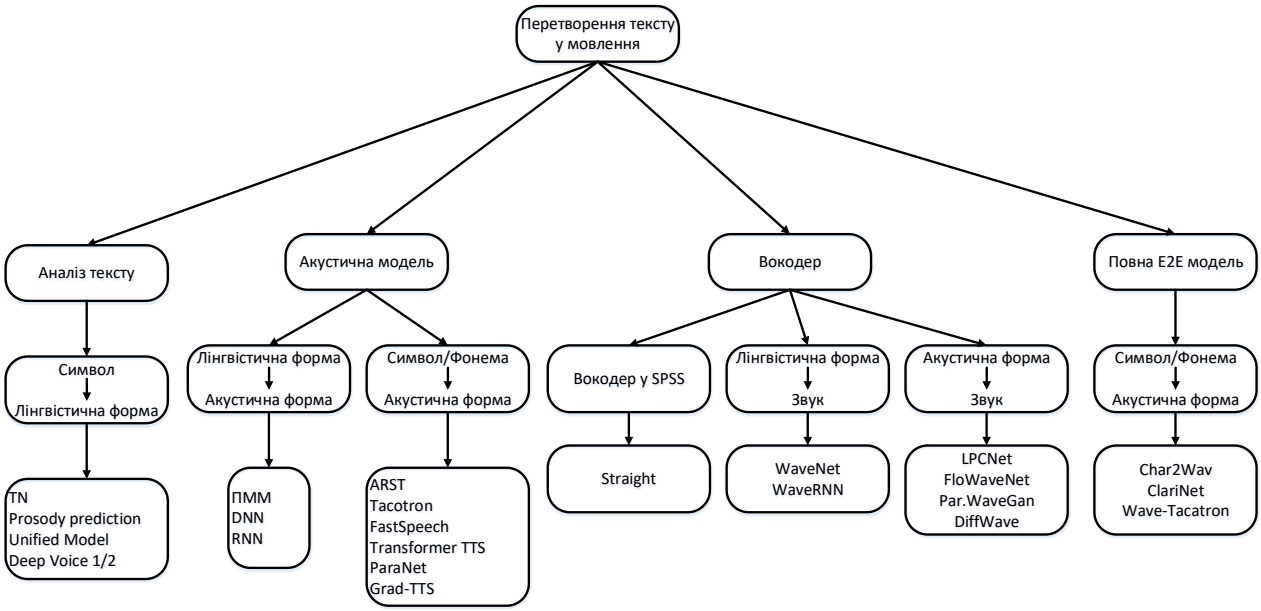

Рис. 1.2. Класифікація компонентів та способів їх реалізації із застосуванням нейронних мереж при перетворенні тексту в аудіо

Така таксономія узгоджується з потоком перетворення даних із тексту в сигнал:

– аналіз тексту перетворює символ у фонему або інші мовні властивості;

– акустичні моделі формують акустичні ознаки або з мовних властивостей, або з символів/фонем;

– вокодери генерують хвилю з мовних або акустичних властивостей;

– повністю наскрізні моделі безпосередньо перетворюють символи/фонеми у хвилю.

Потік даних від тексту до аудіосигналу можна представляти також як показано на рис. 1.3.

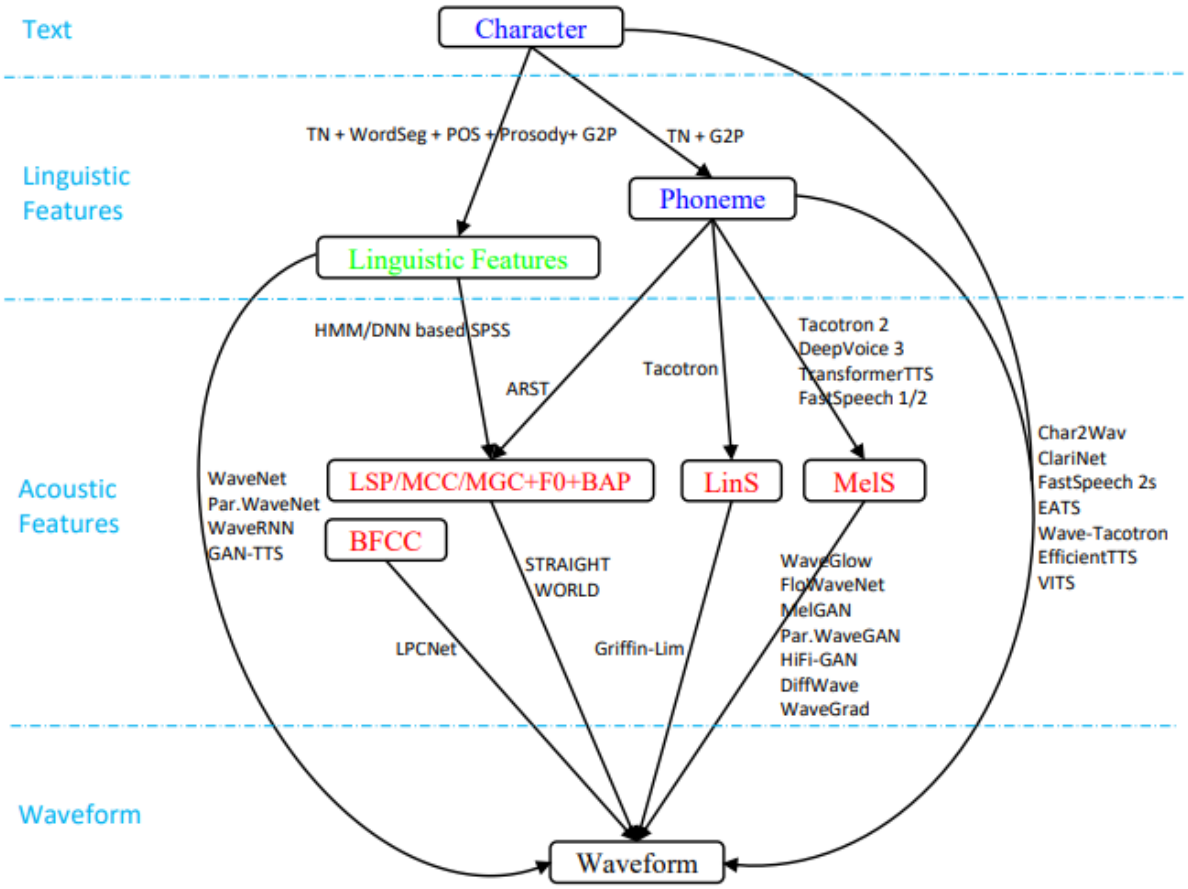

Рис. 1.3. Потік перетворення тексту у голосове повідомлення

У процесі перетворення тексту в мовлення існує декілька представлень даних. Перше – це символьне представлення, яке є необробленим форматом тексту. Наступне представлення – це властивості мови, які одержують шляхом аналізу текстового повідомлення та включають багато контекстної інформації щодо вимови, тембру та інших властивостей. Фонеми є одним із найважливіших елементів лінгвістичних властивостей і, зазвичай, використовуються окремо для представлення тексту в нейронних моделях перетворення тексту в аудіо.

Третє представлення – акустичні властивості, які є абстрактними уявленнями звукового сигналу. При застосуванні статистичного параметричного синтезу голосових повідомлень, LSP (лінійні спектральні пари), MCC (mel-кепстральні коефіцієнти), MGC (mel-узагальнені коефіцієнти), F0 і BAP аналізуються акустичні характеристики, що можуть

бути легко перетворені у форму звукової хвилі за допомогою вокодерів, таких як STRAIGHT і WORLD.

У наскрізних моделях перетворення тексту в аудіо із застосуванням нейронних мереж, мел-спектрограми або лінійні спектрограми, зазвичай, використовуються як акустичні характеристики, які перетворюються на форму сигналу за допомогою нейронних вокодерів.

Waveform, або звукове представлення тексту представляє собою остаточний формат звуку.

Як видно з рис. 2.3, можуть існувати різні потоки даних від вхідного тексту до кінцевого звукового сигналу, включаючи:

1) символ → мовні властивості → акустичні характеристики → сигнал;
2) символ → фонема → акустичні ознаки → хвилеподібна форма;
3) символ → мовні властивості → форма сигналу;
4) символ → фонема → акустичні ознаки → форма хвилі;
5) символ → фонема → сигнал, або символ → сигнал.

1.3. Аналіз текстових повідомлень при перетворені їх в аудіопотік

Аналіз тексту, також відомий як інтерфейс перетворення у TTS, забезпечує трансформацію вхідного текстового повідомлення у лінгвістичні функції, компонентами якої виступають властивості вимови та продосії для спрощення генерації аудіо потоку.

З метою виявлення послідовності векторів, компонентами якого є лінгвістичні ознаки при застосуванні статистичного параметричного синтезу, застосовуються процедури аналізу тексту. Вони включають в себе декілька функцій, зокрема нормалізацію і токенізацію на рівні слів, визначення до якої частини мови належить кожне слово, прогнозування вимови слів, а також трансформацію графем у фонеми.

При безпосередньому перетворенні тексту у звук з використання нейромереж, що забезпечує ефективну здатність моделювання, кожен окремий символ або послідовність фонем безпосередньо виступають у якості вхідних даних генерації звукових повідомлень. Це сприяє зменшенню навантаження при виконанні аналізу тексту. При такому сценарії все ж необхідно проводити нормалізацію текстового повідомлення для одержання стандартизованої форми представлення слів з врахуванням кожного символу, а також доцільно застосовувати перетворення графем у фонеми для отримання їх зі стандартизованого представлення кожного слова.

У деяких працях [6-9] пропонують застосовувати повну наскрізну генерацію, яка прямо синтезує аудіохвилю з вхідних текстових даних, однак проводити процедуру нормалізації тексту все ж варто, оскільки неопрацьований та нестандартизований текст знижує практичність використання такого підходу.

Окрім цього, наявні моделі прямого перетворення тексту у голосові повідомлення містять вбудовані програмні функції для проведення аналізу тексту. Для прикладу, моделі нейронних мереж Char2Wav і DeepVoice 1/2 забезпечують трансформацію букв (символів) у мовні властивості-ознаки безпосередньо у власному конвеєрі, застосовуючи при цьому нейронні мережі. У деяких моделях явно передбачено реалізацію алгоритмів просодії (тембру голосу, темпу і т.п.) за допомогою текстового кодувальника.

Неопрацьоване текстове повідомлення може містити нестандартні слова, які варто перетворити на такі, які вживаються в усному мовленні. Для цього використовують процедури нормалізації, які забезпечують простоту формування вимови. Для прикладу, у тексті зустрічається рік «2023», при його нормалізації і стандартизації одержують значення «дві тисячі двадцять три», «15.09» нормалізовано у «п’ятнадцяте вересня». У ранніх роботах, присвячених процедурам нормалізації тексту, використовувся підхід асоціативних правил, а наступним кроком у конвеєрі було застосування

нейронних мереж. Це давало можливість забезпечувати моделювання нормалізованого текстового повідомлення у вигляді послідовної задачі. У такому випадку послідовність на виході і цільова послідовність були відповідно нестандартизованими токенами і нестандартизованими словами звукового повідомлення.

Для символьних мов, зокрема китайської, проводити токенізацію необхідно для встановлення меж слів у неопрацьованому тексті. Це важливо з точки зору забезпечення точності встановлення приналежності до частини мови, прогнозування вимови та перетворенні графем до фонеми.

Частини мови будь-якого слова у текстовому повідомленні важливі при виконанні трансформації графем у фонеми і при пронозуванні тембру та вимови при перетворенні тексту в аудіо.

Дані про інтонацію, тембр, наголос і ритм відповідають різним тривалостям складу при вимові, силі звуку та висоті, що відображає емоційне забарвлення при спілкуванні людей. При прогнозуванні властивостей вимови тексту використовується система тегів, яка дозволяє позначити кожен з видів просодії. Для різних мов існують різні системи тегування та відповідні інструментальні засоби [14, 16-18]. Прогнозування інтонації та розставлення акцентів на слова для англійської мови можна реалізувати із застосуванням тегів, які імлементовано в інструментальний засіб ToBI. Характерною його особливістю є наявність засобів для формування вимови і різних тривалостей між словами.

Для прикладу, у текстовому повідомленні «Мері пішла в магазин?», «Мері» та «магазин» виступають у ролі слів на які зроблено акцент, а це відповідно супроводжується підвищенням тональності. Синтез голосових повідомлень для китайської мови, зазвичай, містить мітки для розподілу вимови і реалізується:

– для слова – за допомогою токена (PW);
– для фрази – за допомогою токена (PPH);
– для акцентованої фрази (IPH).

На основі таких представлень можна сформувати трирівневе ієрархічне дерево для формування емоційного забарвлення мови.

У деяких дослідженнях [3, 8, 18] при прогнозуванні просодії пропонуються застосовувати різні структурні моделі, зокрема, CRF, RNN, а також трансформери з функцією уваги.

Для спрощення генерації звукових потоків з тексту доцільно проводити перетворення графем у відповідні фонеми. Як приклад графема «speech» трансформується у фонему «сп ій ч». Сформований лексикон відповідності графем та фонем, зазвичай, застосовується при перетворенні текстового повідомлення у голосове. Проте для більшості мов, які є алфавітними, словник відповідності графем та фонем не може забезпечити повноту вимови усіх слів. Отже, конверсія при перетворенні графем у фонеми для алфавітних мов забезпечує синтез вимови слів, що не належать до словникового запасу. Для китайської мови словник відповідності між графемами та фонемами охоплює практично усі ієрогліфи. Проте у цій мові наявно багато паронімів, які ідентифіковуються тільки у контексті символів. Отже, застосування перетворень графеми-фонеми в основному відповідають за усунення багатоголосої неоднозначності, що визначає вимову в контексті конкретного слова.

Після того, як проведено аналіз текстового повідомлення, можна перейти до визначення лінгвістичних властивостей, які є вхідними параметрами наступного етапу конвеєрного опрацювання, наприклад, акустичних моделей у при використання підходу SPSS або вокодерів. Лінгвістичні властивості можна одержати агрегуванням ознак проаналізованого тексту на різних рівнях, включаючи рівні фонеми, склади, слова, фрази та речення в цілому.

1.4. Висновки до розділу

У даному розділі одержано наступні результати:

1 Проведено аналіз підходів і технологій синтезу голосових повідомлень і встановлено, що автоматизація цього процесу потребує вдосконалення процесів аналізу текстової інформації, побудови більш ефективних акустичних моделей та механізмів відтворення звуку.

2 На основі аналізу таксономії процесів перетворення текстових повідомлень в аудіопотік визначено потенційні способи розвитку існуючих нейромережевих моделей, зокрема, в контексті застосування методів машинного навчання для підвищення якості попереднього опрацювання тексту, перетворення графем у фонеми, а також забезпечення можливості їх прогнозування на основі попередньо навчених нейронних моделей.

3 Проаналізовано методи токенізації тексту, визначено основні переваги і недоліки різних видів токенізації, що дало можливість обґрунтувати доцільність застосування архітектури нейронних мереж на основі трансформерів для забезпечення швидкості та ефективності майже в реальному часі.

РОЗДІЛ 2
МОДЕЛІ ТА АЛГОРИТМИ ПОБУДОВИ АКУСТИЧНИХ МОДЕЛЕЙ

2.1. Аналіз та обґрунтування базових характеристик аудіосигналів

Комп’ютер інтерпретує звуковий сигнал, який змінює амплітуду протягом фіксованого періоду часу. Кожна вибірка зазвичай приймає 65536 значень (16 біт), а частота вимірюється в кГц. Приклад форми звукового сигналу показано на рис. 2.1.

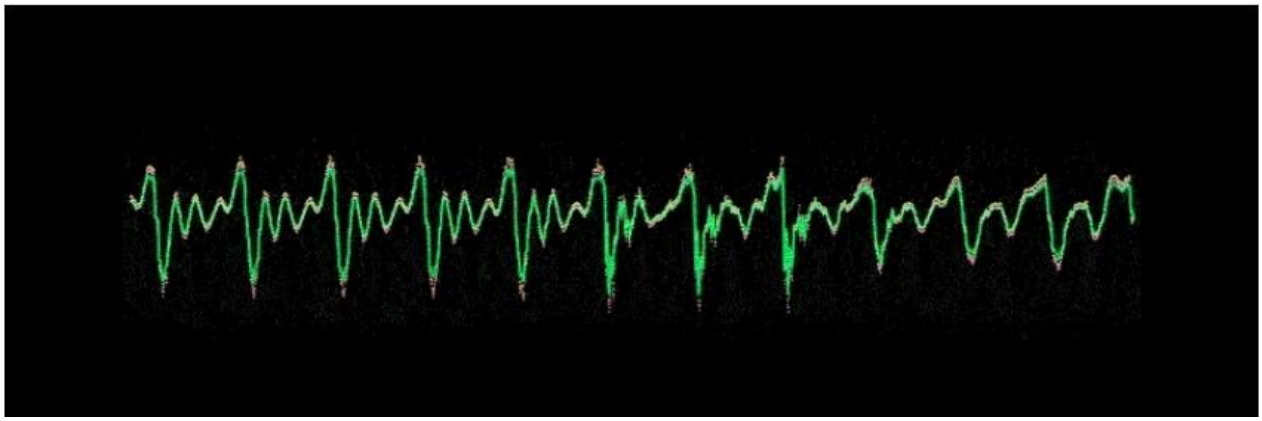

Рис. 2.1. Приклад форми звукового сигналу

Механізм синтезу голосових повідомлень не приймає безпосередньо символи як вхідні дані, а фонеми, зокрема наголошений ARPA (для англійської мови), як це організовано у CMU Dict. Наприклад, зелений колір означає «G R IY1 N». Інші мови можуть мати різні формати, але однакове використання. Ось чому необхідно перетворювати вхідний текст (під час навчання) у фонеми, які є окремими одиницями звуків, які відрізняють одне слово від іншого в мові.

Коли вхідні дані опрацьовуються, вони перетворюються на фонеми за допомогою нейронної мережі, навченої на відомих висловлюваннях, яка також навчається генерувати написання для нових слів.

Моделі глибокого навчання не приймають необроблений аудіо потік безпосередньо як вхідний сигнал, тому аудіо перетворюється на спектрограми, а перетворення Фур’є трансформує вихідний аудіо потік у частотно-часову область. Процес трансформації розбиває тривалість звукового сигналу на менші сигнали перед перетворенням, а потім об’єднує вихідні дані в єдине ціле. Приклад спектрограми показано на рис. 2.2.

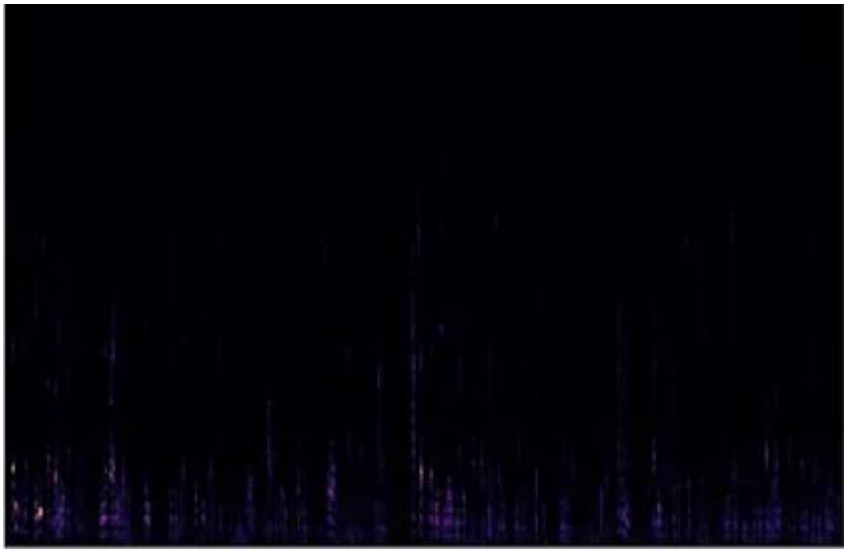

Рис. 2.2. Приклад спектрограми

Колір на рис. 2.2 візуально демонструє звукові децибели, однак з цього рисунку дуже мало інформації про аудіозапис.

Людське розуміння звуку з точки зору частоти може сильно відрізнятися залежно від її зміни, і природа не сприймає звуки лінійно. Це причина, чому була розроблена шкала Мела. Його суть полягає в тому, що він враховує шкалу децибел під час роботи з амплітудами (наскільки голосно) і логарифмічну шкалу для частот (висота звуку). Усі функції голосу зберігаються в спектрограмі Мела, приклад якої представлено на рис. 2.3.

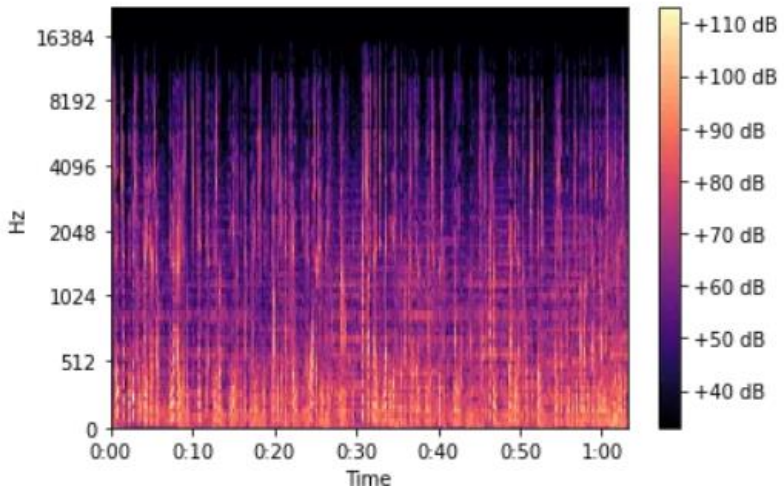

Рис. 2.3. Спектрограма за шкалою Мела

Як видно з рис. 2.3 спектрограма Мела забезпечує набагато чіткішу картину в одиницях вимірювання децибели і є оптимізованою для імплементації при перетворенні тексту в аудіопотік.

Найпоширенішим числовим показником, який інженери ML використовують для оцінки якості синтезу голосових повідомлень, є середня оцінка думки – Mean Opinion Score (MOS), яка коливається від 1 до 5 із повсякденним людським мовленням від 4,5 до 4,8. Для розрахунку MOS використовується наступна формула:

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N} \quad (2.1)$$

де R – індивідуальні оцінки для заданих стимулів N суб'єктів.

На основі авторегресії можна описати модель, яка прогнозує майбутні значення на основі історичних, а в основі ідеї лежить припущення, що кожне наступне значення буде подібним до попереднього. Це відіграє важливу роль у сфері опрацювання аудіо сигналів, оскільки необхідно знати, на наборах слів навчений оратор, щоб отримати правильний вихідний звук.

Під час роботи між цими двома алгоритмами існує також узагальнений компроміс між швидкістю та якістю, де авторегресійна

генерація має нижчу швидкість, але вища якість і неавтоматична регресійна генерація мають зворотний характер.

Статистичний параметричний синтез мовлення (SPSS) – представляє собою підхід Text To Speech для вирішення проблем традиційного конкатенативного перетворення тексту у голосові повідомлення. Цей метод синтезує звук шляхом генерації акустичних параметрів, необхідних для мовлення, а потім відновлення звуку зі згенерованих акустичних параметрів за допомогою алгоритмів.

Один з найвідоміших і поширених фреймворків, заснований на використанні статистичних параметрів синтезу звуку і включає дві основні фази. Загалом, структуру щодо опрацювання та конвеєр перетворення тексту в голосове повідомлення з використанням двох стадій трансформації продемонстровано на рис. 2.4.

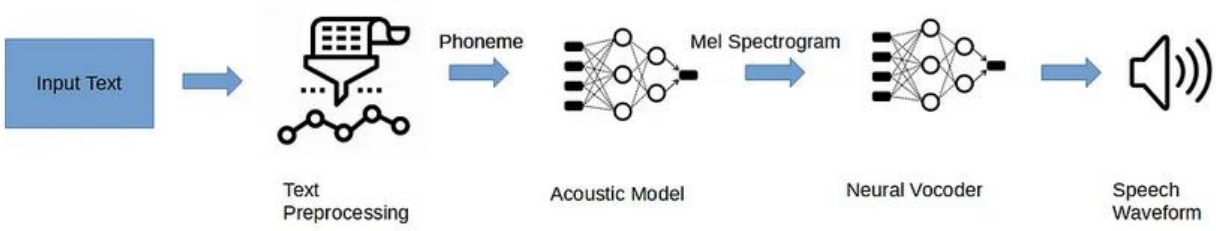

Рис. 2.4. Конвеєр перетворення тексту у голосове повідомлення

В історичному контексті, технології перетворення тексту в аудіопотік виконувались на фреймворках, конвеєрну архітектуру яких показано на рис. 2.5.

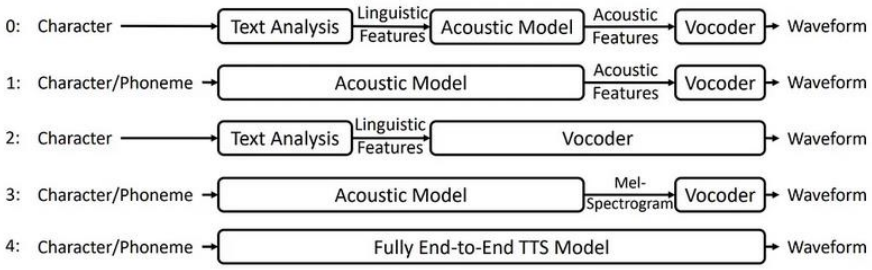

Рис. 2.5. Існуючі фреймворки перетворення тексту в аудіо

Розвиток та еволюція технологій перетворення тексту у голосове повідомлення показано на рис. 2.6.

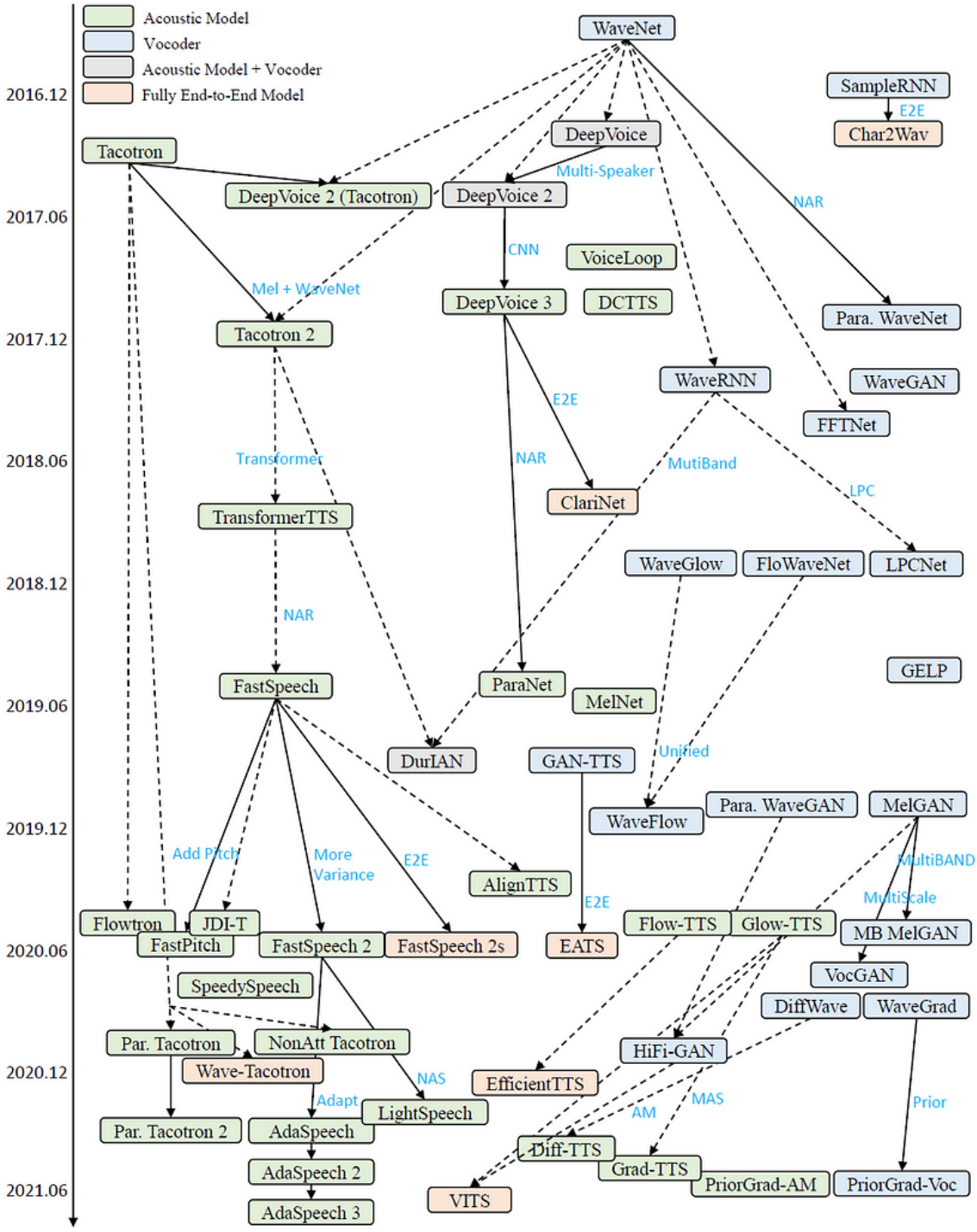

Рис. 2.6. Еволюція розвитку технологій перетворення тексту в аудіопотік

Як видно з рис. 2.6, на сьогодні найбільш ефективним є застосування нейромережевого підходу, як при опрацюванні самого тексту, так і при

синтезі звукових сигналів. Окрім цього, на даному рисунку чітко вказано на те, що сфера синтетичного синтезу мовлення з тексту із використанням нейронних мереж стрімко зростає протягом останніх кількох років і як нещодавно спостерігається тенденція відходу від основної 2-ступеневої (акустичної моделі + вокодер) до прямих моделей наступного покоління.

2.2. Акустичні алгоритми при перетворенні тексту в аудіо

Наразі існують три основні типи архітектури, які використовують поточні акустичні алгоритми. До них належать:

– алгоритми на основі рекурентних нейронних мереж;
– алгоритми на основі згорткових нейронних мереж;
– алгоритми на основі трансформерів.

Рекурентна нейронна мережа може використовуватися для представлення такої структури акустичної моделі та таких алгоритмів, як неавторегресійний Tacotron 2, структуру якого показано на рис. 2.7.

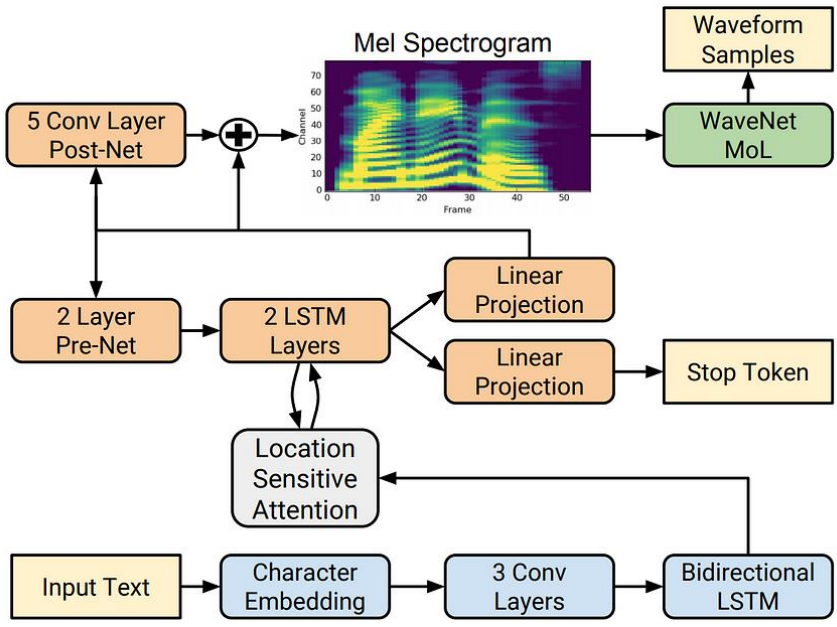

Рис. 2.7. Архітектура Tacotron 2

По факту, RNN мережа виконує прогнозування ознак «sequence-to-sequence», яка в свою чергу забезпечує прогнозування послідовності кадрів Mel-спектрограми з послідовності вхідних символів.

Модифікована версія WaveNet генерує зразки хвилі у часовій області, обумовлені прогнозованими фреймами Mel-спектрограми. Архітектура TacoTron 2 стала величезним кроком вперед у покращенні якості голосу порівняно з іншими методами, такими як конкатенативний, параметричний і авторегресійний TacoTron 1.

Такі алгоритми, як DeepVoice 3, використовують повнозв’язну згорткову структуру мережі для синтезу мовлення, генеруючи Мел-спектрограми з символів і масштабуючи до реальних наборів даних для кількох мовців.

Цей клас акустичних алгоритмів схожий на класичні CNN, які класифікують собак і котів, навчаючись на зображеннях кожного класу, але в цьому випадку навчання відбувається на спектрограмах Мела.

DeepVoice 3 удосконалюється в порівнянні з попередніми системами DeepVoice 1/2, використовуючи більш компактну модель послідовність-послідовність та безпосередньо прогнозуючи спектрограми замість складних лінгвістичних властивостей.

Акустичні алгоритми на основі трансформера (з функцією самоуваги), такі як FastSpeech 1/2, використовують архітектуру енкодер-увага-декодер на основі трансформаторів для генерації спектрограм Мела з фоном. Вони є похідними від мереж трансформації, які є архітектурою, яка спрямована на вирішення послідовних завдань та одночасного легкого опрацювання довгострокових залежностей.

Алгоритми на основі трансформерів значно пришвидшують синтез мовлення через мережу прямого трансформера для паралельного генерування спектрограм Мела на відміну від інших моделей (наприклад, Tacotron 2), які мають авторегресійні декодери уваги енкодера.

Найважливішим є те, що алгоритм спрощує вивід, повністю видаляючи спектрограми Мела як проміжну ланку, і безпосередньо генерує сигнал мовлення з тексту під час логічного виводу, насолоджуючись перевагами повної наскрізної спільної оптимізації під час навчання та низької затримки під час логічного виводу. Архітектуру трансформера для перетворення текстової інформації у голосове повідомлення представлено на рис. 2.8.

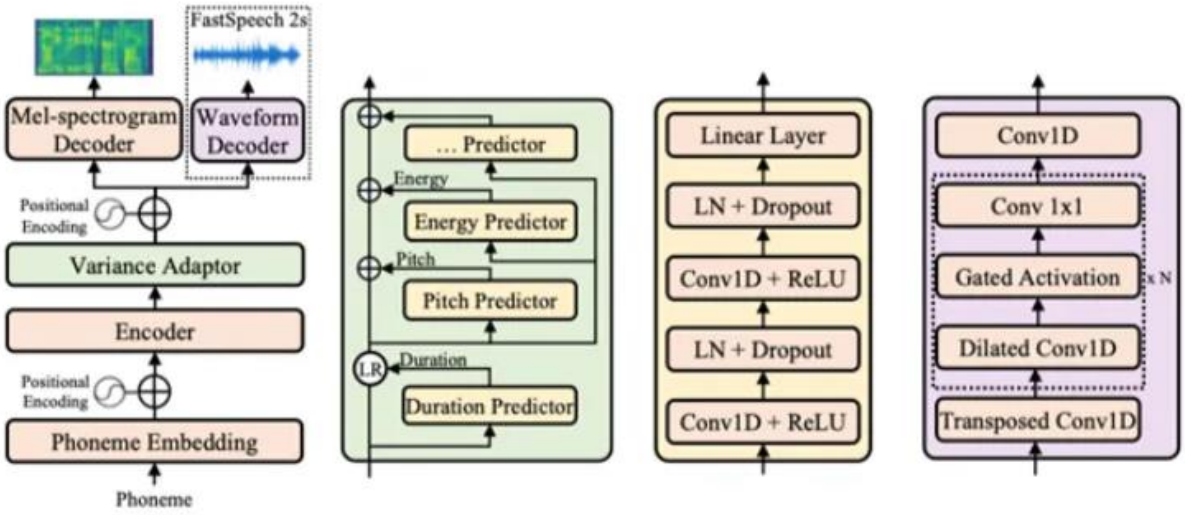

Рис. 2.8. Архітектура FastSpeech на основі трансформерів

FastSpeech 2 забезпечує кращу якість голосу, ніж FastSpeech 1, і зберігає переваги швидкого, надійного та керованого синтезу мовлення завдяки використанню архітектури трансформерів. Потрібно звернути увагу на частину адаптера дисперсії, як на основну відмінність під час використання FastSpeech 2, на відміну від інших акустичних алгоритмів/фреймворків.

2.3. Алгоритми організації та функціонування вокодерів

Вокодери використовуються для перетворення вихідної спектрограми акустичної моделі у цільову звукову форму. Неавторегресійні моделі є

найбільш перспективними, але вони не такі хороші, як авторегресійні моделі.

Існує чотири основних типи вокодерів. Перший тип вокодера використовує авторегресію. WaveNet був першим вокодером на основі нейронної мережі, який розширює згортки таким чином, щоб генерувати точки форми сигналу авторегресією.

Нова частина цього підходу полягає в його здатності майже не використовувати попереднього розуміння вхідних аудіосигналів і натомість покладається на наскрізне навчання.

Другий тип вокодера – на основі потоку. Цей тип організований за допомогою генеративної моделі, яка дає можливість перетворити щільність розподілу ймовірностей на основі послідовного зворотного відображення. Цей підхід реалізується у двох різних системах: одна використовує авторегресійні перетворення, а інша – двосторонні перетворення.

Третій тип вокодера організовується на основі GAN, який побудовано на основі типових генеративних жадних мереж (GAN), які використовуються для задач формування зображень. Будь-яка структура GAN складається з генератора для синтезу даних і дискримінатора для визначення даних генератора. У відповідності до цього виникає аналогія про грабіжники та поліцейського, які завжди намагаються придумати нові схеми/напади, а поліцейські намагаються їм перешкодити.

Більшість поточних вокодерів на основі GAN використовуватимуть розширену згортку для збільшення сприйнятливого поля для моделювання тривалої залежності в послідовності сигналу та транспоновану згортку для підвищення дискретизації інформації про стан відповідно до довжини сигналу.

З іншого боку, дискримінатори зосереджуються на розробці моделей для захоплення характеристик сигналу, щоб забезпечити кращий керівний сигнал для генератора. Функція втрат покращує стабільність і ефективність змагального навчання та покращує якість звуку. Як видно з наведеної нижче

табл. 2.1, багато сучасних нейронних вокодерів базуються на GAN і використовуватимуть різні підходи з функціями генератора, дискримінатора та функцією втрат.

Таблиця 2.1

Реалізації підходу GAN

Підхід GAN	Генератор	Дискримінотор	Функція втрат
WaveGan	DCGAN	-	WGAN-GP
GAN-TTS	-	Random Window D	Hinge_Loss GAN
MelGan	-	Multi-Scale D	LS-GAN Feature Matching Loss
Par. WaveGAN	WaveNet	-	LS-GAN Multi-STFT Loss
HiFi-GAN	Multi-Receptive Field Fusion	Multi-Period D, Multi_Scale D	LS-GAN, STFT Loss, Feature Matching Loss
VocGAN	Multi-Scale G	Hierarchical D	LS-GAN, STFT Loss, Feature Matching Loss
GED	-	Random Window D	Hinge-Loss GAN, Repulsive Loss

Інший тип алгоритмів організації вокодерів базується на основі дифузії: використання ймовірнісних моделей дифузії з усуненням шуму для вокодерів з інтуїтивним уявленням про те, що відображення між даними та латентними розподілами з процесом дифузії та зворотним процесом.

Сучасні дослідження показують, що цей клас вокодерів створює високоякісні голосові повідомлення, але потребує значної кількості часу для

формування виводу. У табл. 2.2. – 2.4 представлено усі види вокодерів на основі нейронних мереж з відповідними їм архітектурами.

Таблиця 2.2

Нейромережеві вокодери без моделей

Вокодер	Вхідні дані	Авторегресійна (А)/ Неавторегресійна модель (НА)	Архітектура
WaveNet	Лінгвістичні ознаки	А	CNN
Sample RNN	-	А	RNN
WaveRNN	Лінгвістичні ознаки	А	RNN
LPCNet	BFCC	А	RNN
Univ. WaveRNN	Спектрограма Мела	А	RNN
SC-WaveRNN	Спектрограма Мела	А	RNN
MB WaveRNN	Спектрограма Мела	А	RNN
FFTNet	Кепстр	А	RNN

У табл. 2.3 представлено реалізації нейронних мереж в основі яких лежать моделі потоків даних. Більшість з них базуються на не авторегресійних моделях перетворення тексту у звук.

Таблиця 2.3

Нейромережеві вокодери на основі моделей потоків даних

Вокодер	Вхідні дані	Авторегресійна (А)/ Неавторегресійна модель (НА)	Архітектура
Par. WaveNet	Лінгвістичні ознаки	НА	
WaveGlow	Спектрограма Мела	НА	Гібридна/CNN
FloWaveNet	Спектрограма Мела	НА	Гібридна/CNN

Продовження табл. 2.3

Вокодер	Вхідні дані	Авторегресійна (А)/ Неавторегресійна модель (НА)	Архітектура
WaveFlow	Спектрограма Мела	А	Гібридна/ CNN
SqueezeWave	Спектрограма Мела	НА	CNN

Наступний клас нейромереж, що використовуються при перетворенні текстових повідомлень у текст є моделі, в основі яких лежать генеративні алгоритми. Коротка характеристика цих моделей представлена у вишляді табл. 2.4.

Таблиця 2.4

Характеристики нейромереж з генеративними алшгоритмами

Вокодер	Вхідні дані	Авторегресійна (А)/ Неавторегресійна модель (НА)	Архітектура
WaveGAN	-	НА	CNN
GELP	Спектрограма Мела	НА	CNN
GAN-TTS	Лінгвістичні ознаки	НА	CNN
MelGan	Спектрограма Мела	НА	CNN
Par. WaveGan	Спектрограма Мела	НА	CNN
HiFi-GAN	Спектрограма Мела	НА	Гібридна/CNN
VocGan	Спектрограма Мела	НА	CNN
GED	Лінгвістичні ознаки	НА	CNN
Fre-GAN	Спектрограма Мела	НА	CNN

Характеристики моделей нейронних мереж, які в своїй основі використовують природу дифузії, представлено у табл. 2.5.

Таблиця 2.5

Типи вокодерів на основі дифузійних моделей

Вокодер	Вхідні дані	Авторегресійна (А)/ Неавторегресійна модель (НА)	Архітектура
WaveGrad	Спектрограма Мела	НА	Гібридна/CNN
DiffWave	Спектрограма Мела	НА	Гібридна/CNN
PriorGrad	Спектрограма Мела	НА	Гібридна/CNN

Новим трендом при вдосконаленні методів і моделей перетворення тексту у голосові повідомлення є наскрізне перетворення. Конвеєр цього перетворення матиме вигляд, як показано на рис.2.9.

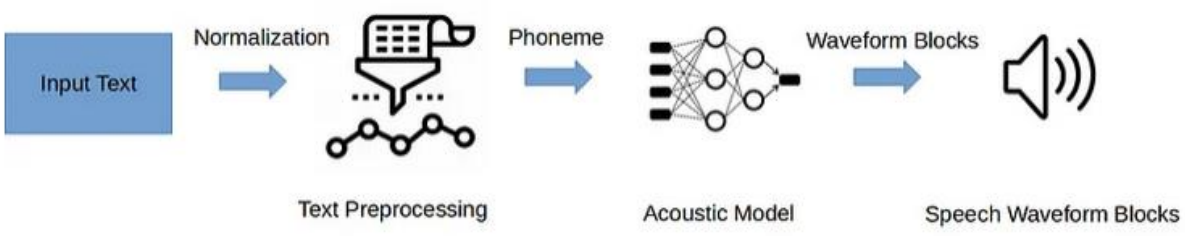

Рис. 2.9. Конвеєр наскрізного перетворення тексту у звук

Поточні дослідження та розвиток технологій перетворення тексту в аудіопотік, як видно з рис. 2.8, рухаються до наскрізного TTS, хоча він ще не досяг критичної маси через деякі поточні обмеження якості та вимоги до часу навчання порівняно з основними двостадійними методами.

Зважаючи на це, наскрізний TTS має деякі явні переваги порівняно з SPSS: Значно знижує витрати на розробку та розгортання. Наявність архітектури об’єднання та наскрізної оптимізації означає зменшення

поширення помилок у поточному основному 2-етапному методі. Вимагає менше загальної анотації та розробки функцій. Традиційні акустичні моделі потребують узгодження лінгвістичних та акустичних ознак, на відміну від нейромережевих моделей типу «послідовність-послідовність», які неявно досліджують узгодженість із застосуванням функції уваги або алгоритмів прогнозування. Такий підхід забезпечує більш пряме перетворення тексту у голосове повідомлення і практично не потребує попереднього опрацювання. Із зростанням потужності моделей на основі нейронних мереж, лінгвістичні ознаки стають більш простими у послідовностях символів або фонем, а акустичні властивості стають більш комплексними.

2.4. Архітектура нейронної мережі для перетворення тексту в аудіопотік

З точки зору генерації природного звучання голосу, нейронні системи перетворення тексту в аудіопотік домінують у сфері мистецтва за шкалою MOS.

Ці нейронні моделі TTS розроблені на базі таких процесорів, як GPU або TPU. Невелика увага приділяється дослідженню можливостей досягнення автономного функціонування моделі на пристрої. Зокрема, авторегресійні моделі, такі як Tacotron2, Deep Voice 3 і Transformer TTS, за своєю суттю повільні. У той час як неавторегресійні нейронні мережі перетворення тексту в аудіопотік FastSpeech2 і Mixer-TTS, є швидкими та мають конкурентоспроможну якість голосу, яку можна порівняти з авторегресійними аналогами. Проте ці моделі мають великі розміри, що робить їх непридатними для застосування на периферійних пристроях з обмеженою пам’яттю.

Нещодавні спроби створити нейронні мережі TTS, які б функціонували на IoT пристроях, включають On-device TTS, LiteTTS, PortaSpeech, LightSpeech і Nix-TTS.

Моделі перетворення тексту в аудіо, які працюють на пристроях по типу Raspberry PI або Arduino є повільним і ресурсомістким, оскільки вони представляють собою модифіковані версії Tacotron2 для формування спектрограми Мела і використовують WaveRNN для вокодера.

Хоча LiteTTS може генерувати звук з тексту, він все ще вимагає значних ресурсів, зокрема, 13,4 млн параметрів. Крім того, двоступінчасті моделі перетворення тексту в голос все ж кращі в плані як навчальної стабільності, так і синтезу якісного звуку.

PortaSpeech використовує моделі VAE і Flow для створення спектрограми Мела. Найменша версія має параметри 6.7M і характеризується помітним погіршенням якості голосу.

LightSpeech використовує пошук нейронної архітектури (NAS) для зменшення розміру моделі FastSpeech2. Хоча отримана у результаті модель має 1,8 млн. параметрів, процес NAS, як відомо, вимагає інтенсивних обчислень і має величезний вплив на навколишнє середовище.

Крім того, NAS чутливий до модифікації. Архітектура моделі, оптимізована для одного мовного набору даних (наприклад, англійської), не гарантовано працюватиме на іншому (наприклад, корейському).

NixTTS застосував дистиляцію знань, щоб зменшити розмір VITS до 5,2 МБ шляхом окремого навчання текстового енкодера в латентний енкодер і декодера латентного сигналу у форму хвилі. Незважаючи на значне зменшення розміру, декодер є одноразовим або спеціальним для енкодера, на відміну від вокодерів загального призначення, таких як HiFiGAN, який доступний у моделі з суб-параметрами 1M для периферійних пристроїв.

Незважаючи на те, що згадані вище моделі сприяють ефективності перетворення тексту у звук на міні комп’ютерах, перевірка на процесорах ARM не проводилась, за винятком Nix-TTS, який використовував скомпільовану модель ONNX. Крім того, більшість із цих моделей не мають загальнодоступних реалізацій.

Таким чином, відтворюваність, справедливе порівняння і аналіз важко виконати. У роботі пропонується модель EfficientSpeech для перетворення тексту в аудіопотік з природним звучанням, яка підходить для периферійних пристроїв.

EfficientSpeech використовує неглибоку піраміду U-Network трансформерного енкодера фонеми та неглибокий транспонований згортковий блок як декодер спектрограми Мела. Дана модель має лише 266 тис. параметрів, приблизно 15% від розміру LightSpeech або 0,8% від FastSpeech2. EfficientSpeech споживає 90 MFLOPS лише для створення 6 секунд мел-спектрограми. Використовуючи компактну версію HiFiGAN [13], загальні параметри моделі становлять 1,2 МБ або 22% сигналу тексту в мову Nix-TTS.

Використовуючи HiFiGAN як вокодер, він працює з RTF 1,7 для генерації голосу на Raspberry Pi 4. Без накладних витрат вокодера генерація мел-спектрограми відбувається зі швидкістю RTF 104,3.

EfficientSpeech досягає конкурентоспроможної CMOS на рівні -0,14 під час навчання на наборі даних LJSpeech і оцінки за допомогою FastSpeech2. Завдяки невеликому розміру EfficientSpeech можна навчити на одному GPU за 12 годин. Архітектура EfficientSpeech показана на рис. 2.10.

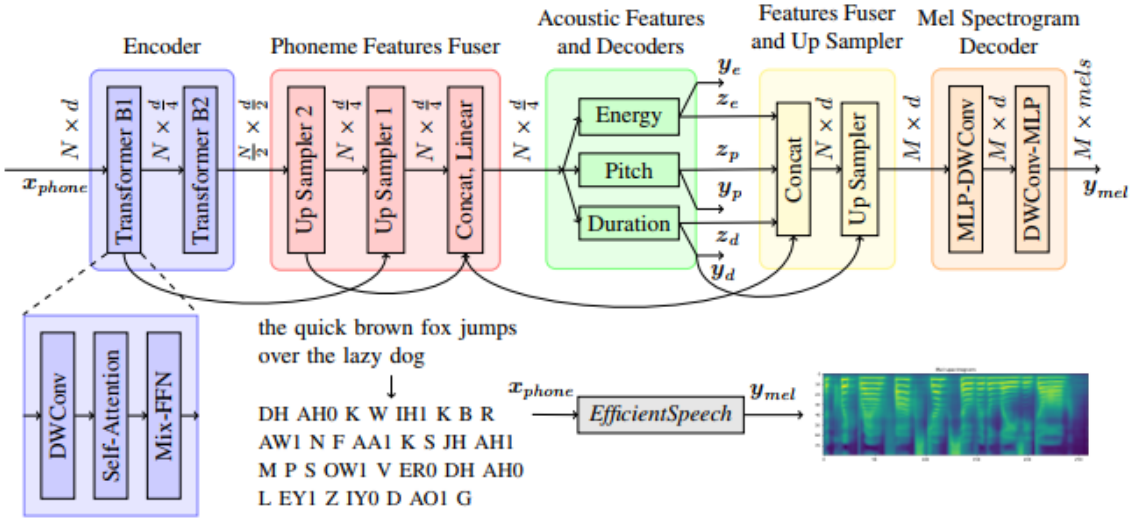

Рис. 2.10. Архітектура моделі EfficientSpeech.

Енкодер фонеми складається з двох блоків трансформерів, об’єднаних функціями дискретизації, що нагадує U-Net. EfficientSpeech використовує паралельні акустичні функції та виконує прогнозування. Акустичні характеристики об’єднуються з функціями фонем та дискретизуються для декодування мел-спектрограми, яка складається з двох блоків.

Послідовність фонем $xphone \in \mathbb{R}^{N \times d}$ є векторним представленням (embedding) вхідного тексту. Усі згорткові шари є одновимірними. N – змінна довжини послідовності фонем, тоді як $d = 128$ — розмір вектора при застосуванні процедури embedding.

Енкодер фонем складається з 2-х блоків трансформерів. Кожен блок складається з роздільної по глибині згортки для об’єднання самоуваги між об’єднаними функціями та Mix-FFN для добування нелінійних функцій.

Mix-FFN подібний до типового трансформера FFN, за винятком додаткового шару згортки та використання активації GeLU між двома лінійними шарами.

Нормалізація шару застосовується після функції самоуваги і Mix-FFN. І функція самоуваги, і Mix-FFN використовують залишкове з’єднання для швидкої конвергенції.

Перший блок трансформера зберігає довжину послідовності, одночасно зменшуючи розмірність ознак на чверть. Другий трансформер зменшує довжину послідовності вдвічі, одночасно подвоюючи розмірність ознак.

Кожна вихідна функція блоку трансформера збільшується за допомогою лінійного шару та транспонованого згорткового шару. Рівень ідентичності замінює транспоновану згортку, якщо цільова форма об’єкта $N \times \frac{d}{4}$ вже одержана. Потім обидві функції зливаються разом, щоб сформувати остаточні функції фонеми. Цей стиль архітектури U-Network був наслідуваний з SegFormer, що використовується для семантичної сегментації в задачах комп'ютерного зору.

Зменшення розмірності функції та довжини послідовності знижує FLOPS і кількість параметрів моделі. Блок акустичних функцій і декодерів запозичив ідею з дисперсійного адаптера FastSpeech2. Це змушує мережу прогнозувати енергію y_e, висоту звуку y_p і тривалість y_d.

Різниця у пропонованій реалізації полягає в тому, що замість прогнозування акустичних параметрів послідовно, EfficientSpeech генерує їх паралельно, що призводить до швидшого формування відповіді.

Прогнозовані значення y_e, y_p і y_d генеруються двома блоками «Conv-LN-ReLU» та вихідним лінійним шаром (з ReLU для тривалості для забезпечення позитивних значень).

Об’єднані функції енергії та висоти вбудовуються на останньому рівні для створення енергії z_e та висоти z_p. Тим часом тривалість z_d добувається перед активацією ReLU.

У блоці властивостей «Fuser» і «Up Sampler» усі акустичні функції повторно використовуються та зливаються разом із функціями фонем. Потім об’єднані об’єкти дискретизуються до правильної довжини послідовності спектрограми Мела M з використанням прогнозованої тривалості.

Останнім структурним елементом архітектури є декодера спектрограми Мела. До його складу входить 2 блоки лінійного шару і два шари глибокої роздільної згортки. Кожен шар використовує активацію функції тангенса, а потім LN.

2.5. Висновки до розділу

Основні результати даного розділу полягають в наступному:

1. Визначено основні характеристики аудіосигналів, які можна ефективно використати при перетворенні тексту у звук, що дало змогу врахувати, зокрема, спектрограми Мела при побудові компонентів архітектури нейронної мережі перетворення тексту в аудіопотік.

2. Досліджено алгоритми побудови вокодерів, які використовуються для перетворення вихідної спектрограми акустичної моделі у цільову звукову форму, що дало змогу обґрунтувати застосування GAN вокодерів при проектуванні архітектури нейронної мережі перетворення текстової інформації в аудіопотік і розпаралелити його.

3. Запропоновано архітектуру нейронної мережі до складу якої входить енкодер на базі трансформерів, які забезпечують зменшення розмірності вхідної матриці фонем у 4 рази та добувають фонетичні властивості, а також декодера, який сформований з блоків визначення акустичних властивостей, зокрема енергії, тривалості і висоти звуку, та блоків декодування властивостей аудіосигналу на основі якого відбувається генерація спектрограми Мела, що дало змогу забезпечити високу продуктивність і якість формування голосового повідомлення у порівнянні з іншими моделями.

РОЗДІЛ 3
ІМПЛЕМЕНТАЦІЯ КОМП'ЮТЕРНОЇ СИСТЕМИ ПЕРЕТВОРЕННЯ ТЕКСТУ В АУДІОПОТІК НА ОСНОВІ RASPBERRY PI

У роботі пропонується реалізувати комп'ютерну систему перетворення тексту в аудіопотік на апаратному забезпеченні мінікомп'ютера Raspberry PI. При такому підході служба аудіомовлення повинна працювати як однорангова програма на базі PubNub Data Stream. Виходячи з цього, з однієї сторони повинен бути наявний одноранговий клієнт-запитувач, який надсилає запит на трансляцію, а з іншого боку, на базі Raspberry Pi повинно працювати програмне забезпечення трансляції мовлення. Клвєнт надсилає текстове повідомлення в межах корисного навантаження PubNub, а транслятор перетворює його в аудіопотік та надсилає на аудіовихід Raspberry Pi.

3.1. Організація схеми підключення пристроїв комп'ютерної системи перетворення тексту в аудіопотік

На початку практичної імплементації проекту потрібно налаштувати аудіодрайвер для Raspberry Pi. Для цього в якості апаратної складової використовується стандартна аудіо плата (рис. 3.1), яка підключається до одного з портів USB, а в якості пристрою виводу звуку може застосовуватися, наприклад, стандартний настільний динамік від Lenovo (рис. 3.2). Проте може бути використана будь-яка акустична система.

Рис. 3.1. USB аудіо плата

Рис. 3.2. Акустична система Lenovo

Схема з’єднання компонентів для трансляції аудіопотоку з текстових повідомлень на апаратному рівні представлено на рис. 3.3.

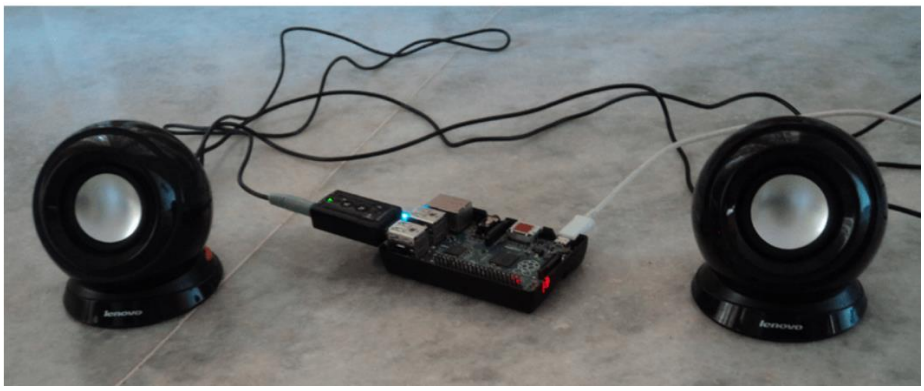

Рис. 3.3. Схема підключення компонентів системи відтворення аудіопотоку

Далі потрібно налаштувати апаратне забезпечення Raspberry Pi для правильного відтворення звукових сигналів.

3.2. Налаштування аудіопристроїв на Raspberry PI

Операційна система Raspbian наслідує Advanced Linux Sound Architecture (ALSA) для керування аудіопристроями. Для налаштування пристроїв відтворення звуку потрібно встановити кілька пакетів, щоб перевірити звуковий пристрій через ALSA. За допомогою утиліти apt-get

потрібно інсталювати наступні пакети під профілем адміністратора. На рис. 3.4 наведено команди для інсталяції необхідних пакетів для роботи з аудіо на Raspberry PI.

```
apt-get install alsa-utils
```

```
apt-get install mpg321
```

```
apt-get install lame
```

Рис. 3.4. Команди встановлення пакетів для роботи з аудіо на Raspberry PI

Перша команда на рис. 3.4 призначена для встановлення пакету Alsa-utils. Цей пакет містить налаштування параметрів, які показані на рис. 3.5.

```
amixer                      - консольний мікшер
alsamixer                   - інтерактивний мікшер
amidi                       - читання і запис в ALSA RawMIDI порт
aplay, arecord              - консольне відтворення і запис звуку
aplaymidi, arecordmidi      - консольне відтворення і запис MIDI
aconnect, aseqnet, aseqdump - консольне MIDI sequencer contro
```

Рис. 3.5. Вміст пакету Alsa-utils

Друга команда на рис. 3.4 забезпечує встановлення програвача MP3 файлів для Unix-подібних операційних систем, що забезпечують підтримку виводу звуку через ALSA. Остання команда – встановлення енкодері для MP3 типів файлів. Для коректної роботи Raspberry PI при виводі аудіопотоку необіхдно завантажити драйвер для роботи зі звуком. Для цього використовується команда, показана на рис. 3.6.

```
modprobe snd-bcm2835
```

Рис. 3.6. Завантаження аудіодрайвера для Raspberry PI

Подальше налаштування системи відтворення звукових сигналів передбачає виконання наступних кроків:

– вимкнути мінікомп’ютер Raspberry PI;
– підключити звукову карту USB до одного з USB-портів PI;
– увімкнути Raspberry Pi.

Після того, як відбулось завантаження ОС Raspbian необхідно переконатися, що звукове обладнання виявлено. Щоб перевірити це, потрібно увійти до терміналу (через SSH або програму LXTerminal) і ввести команду «lsusb».

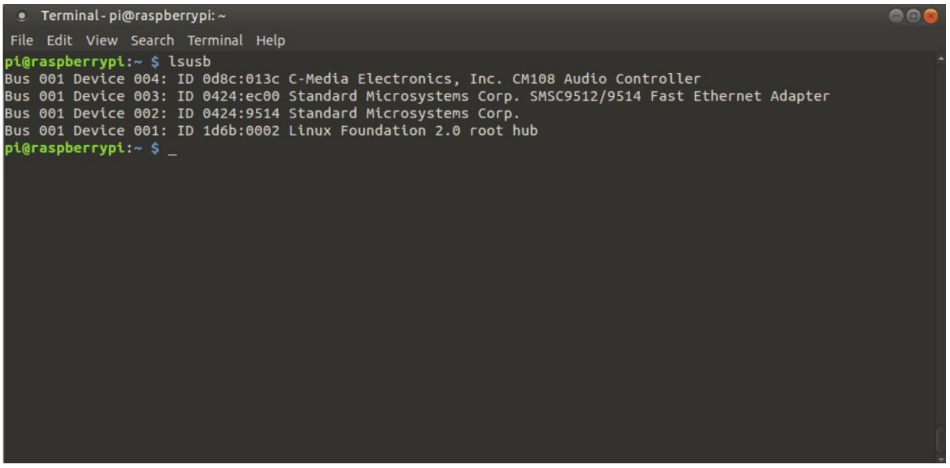

Рис. 3.7. Результат виконання команди «lsusb»

Як видно з рис. 3.7, першим пристроєм у списку портів USB є звукова аудіокарта. Наступний крок полягає в увімкненні аудіопристрою в ALSA.

За замовчуванням звуковий драйвер Raspberry Pi налаштовано на використання вбудованого аудіопристрою PCM. Оскільки використовується зовнішня USB аудіокарта, потрібно виконати деякі налаштування конфігурації, щоб повідомити ALSA про зовнішній аудіо пристрій. Для початку варто перевірити, у якому порядку були завантажені аудіокартки (рис. 3.8).

```
$ cat /proc/asound/modules
 0 snd_bcm2835
 1 snd_usb_audio
```

Рис. 3.8. Перевірка наявних аудіоплат

Відображення доступнх аудіопристроїв починається з 0, тому bcm2835 за замовчуванням завантажується першим, а USB-карта, до якої підключено динамік, завантажується другою. У випадку, якщо потрібно безпосередньо використовувати звичайний аудіороз'єм (без USB) без будь-яких конфігурацій, то можна напряму під'єднати до нього динаміки та відтворювати будь-яке аудіо через карту аудіопристрою PCM, яка використовується за замовчуванням.

Щоб змінити порядок аудіокарток, спочатку створюється файл під назвою «/etc/modprobe.d/alsa-base.conf». Його можна назвати як завгодно, якщо він закінчується на .conf. Потім необхідно додати програмний код, який представлено на рис. 3.9.

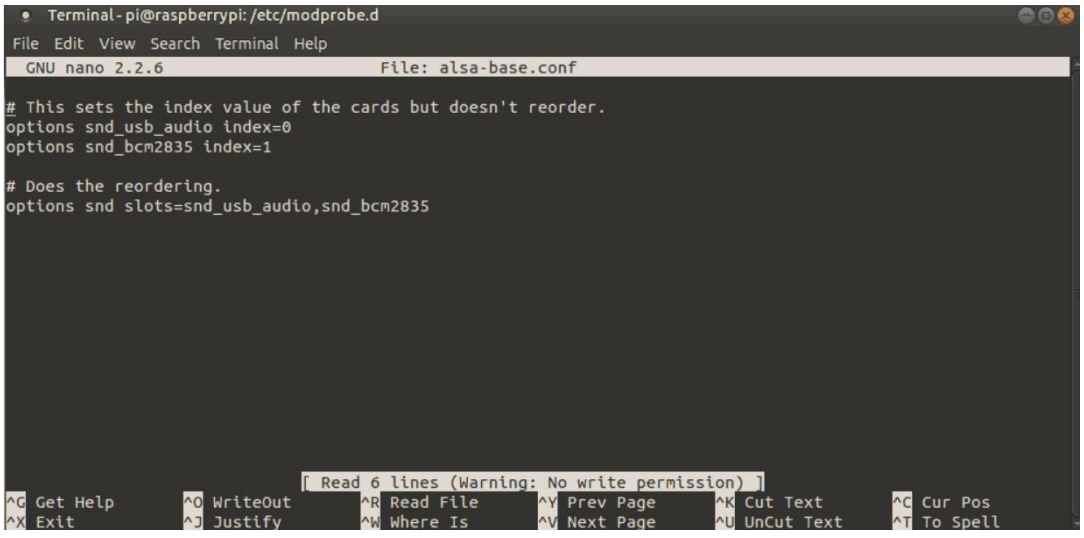

Рис. 3.9. Внесення змін для вибору зовнішньої аудіокарти

Правильним підходом є написання коментарів для пояснення того, що робить кожен рядок програмного коду. Як тільки виконається налаштування

показане на рис. 3.9, потрібно перезавантажити машину, щоб застосувати зміни. Після перезавантаження буде забезпечена можливість відтворення будь-якого аудіо через аудіокартку USB, а сама карта буде використосуватися за замовчуванням, тобто індекс=0. Правильність налаштування і пріоритет апаратного забезпечення для відтворення звуку показано на рис. 3.10.

```
0 snd_usb_audio
1 snd_bcm2835
```

Рис. 3.10. Правильність налаштування зовнішніх пристроїв відтворення звуку

У випадку, коли потрібно повернутися до конфігурації PCM за замовчуванням необхідно відредагувати файл «/etc/modprobe.d/alsa-base.conf», змінивши значення індексу та порядок слотів.

Перевірка і тестування аудіопристроїв відбувається після перезапуску Raspberry Pi шляхом виконання команди «$ amixer» з підказки терміналу, щоб побачити поточну конфігурацію звуку (рис. 3.11).

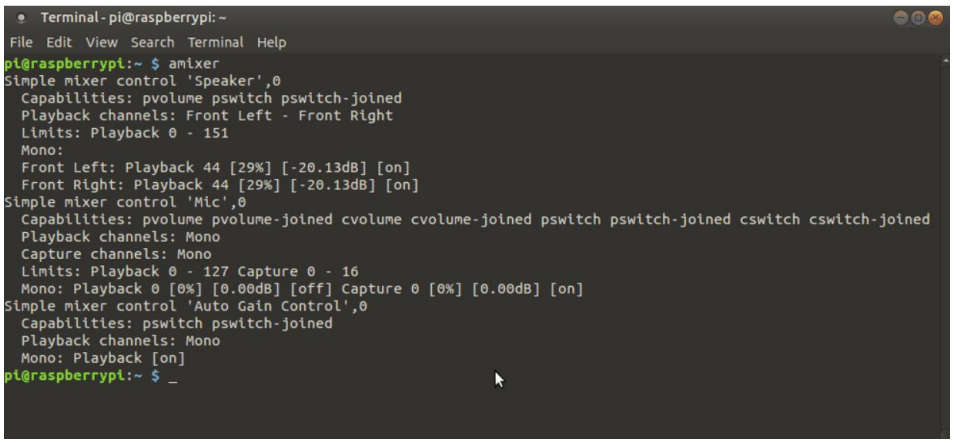

Рис. 3.11. Поточна конфігурація пристроїв відтворення звуку

Як можна помітити з рис. 3.11, у конфігурації відображається детальна інформація про можливості звукової карти, включаючи відтворення та захоплення звуку.

Далі потрібно обрати аудіовихід, який використовуватиме Raspberry Pi. Для цього потрбіно відкрити інструмент налаштування «raspi-config» з командного рядка та виконати наведену нижче процедуру:

– обрати «Послідовність меню» (рис.3.12);
– обрати «Додаткові параметри» (рис. 3.13);
– вибрати «Аудіо» та обрати «Авто» (рис. 3.14).

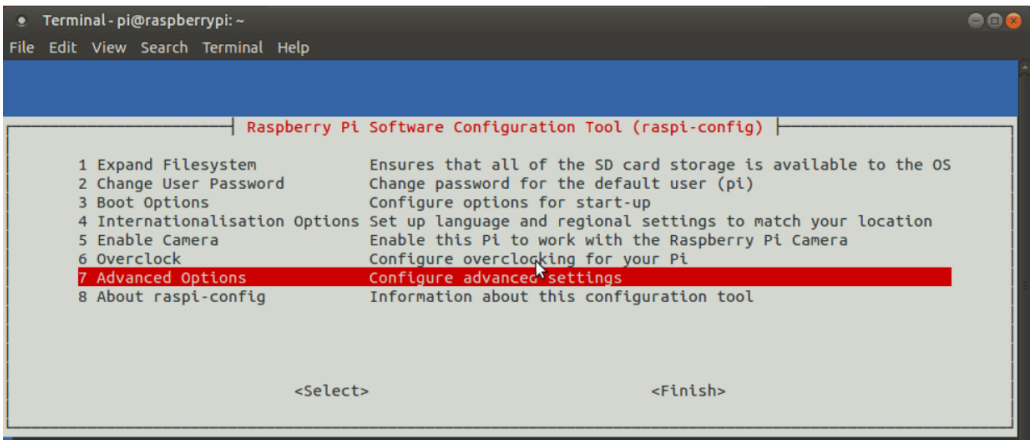

Рис. 3.12. Меню для вибору розширених налаштувань

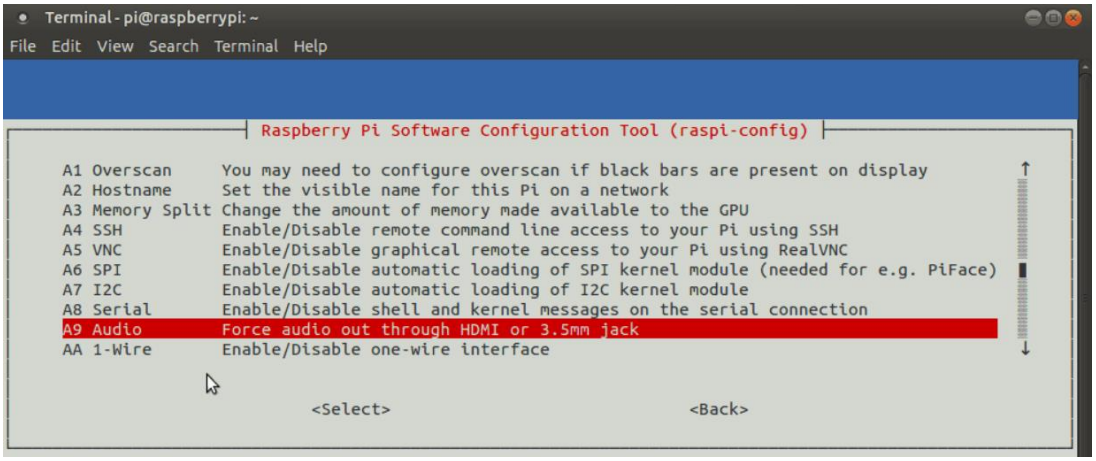

Рис. 3.13. Вибір налаштування порту для відтворення аудіо

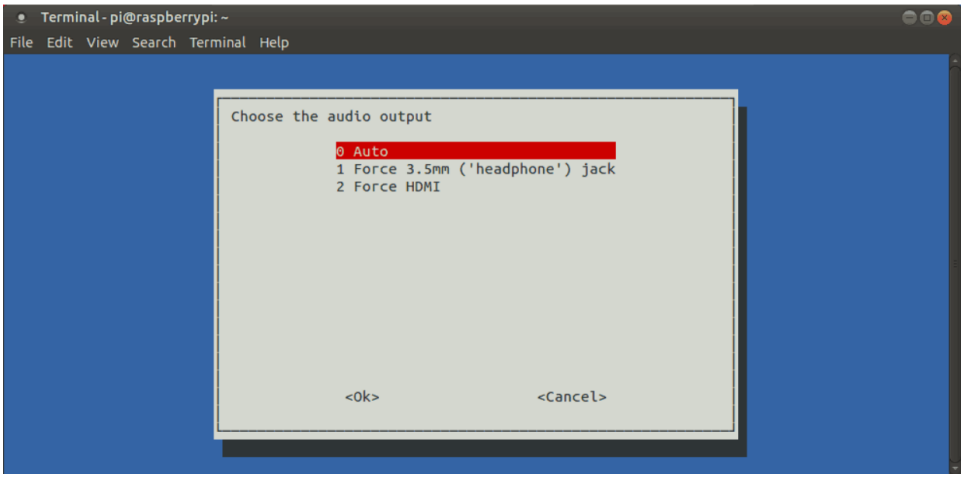

Рис. 3.14. Вибір автоматичного режиму відтворення аудіопотоку

Після вибору режиму «Авто» натискається відповідна клавіша для вибору опції та вийти зі списку параметрів. Це дозволить Raspberry Pi автоматично визначати звуковий вихід і використовувати або HDMI, або аудіороз'єм залежно від того, який з них підключено.

Налаштувавши пристрій відтворення звуку, далі потрібно протестувати настройки. Для цього підключаються динаміки до відповідного виходу на звуковій карті та відтворюється один із тестових звукових файлів, які наявні в операційній системі Raspbian OS. Команда для запуску тестового аудіофайлу і відповідний результат показано на рис. 3.15.

```
aplay /usr/share/scratch/Media/Sounds/Vocals/Singer1.wav
```

Рис. 3.15. Тестування відтворення аудіофайлу

Якщо все пройшло успішно, то буде відтворений швидкий фрагмент голосу оперного співака. Якщо гучність звуку занадто висока або низька, то її можна відрегулювати за допомогою утиліти «alsamixer». Крім того, можна виконати наведені нижче дії, щоб перевірити, чи аудіо працює коректно. Для цього потрібно застосувати утиліту для перевірки динаміків, щоб відтворити синусоїду через порт USB (рис. 3.16).

```
speaker-test -t sine -f 440 -c 2 -s 1
```

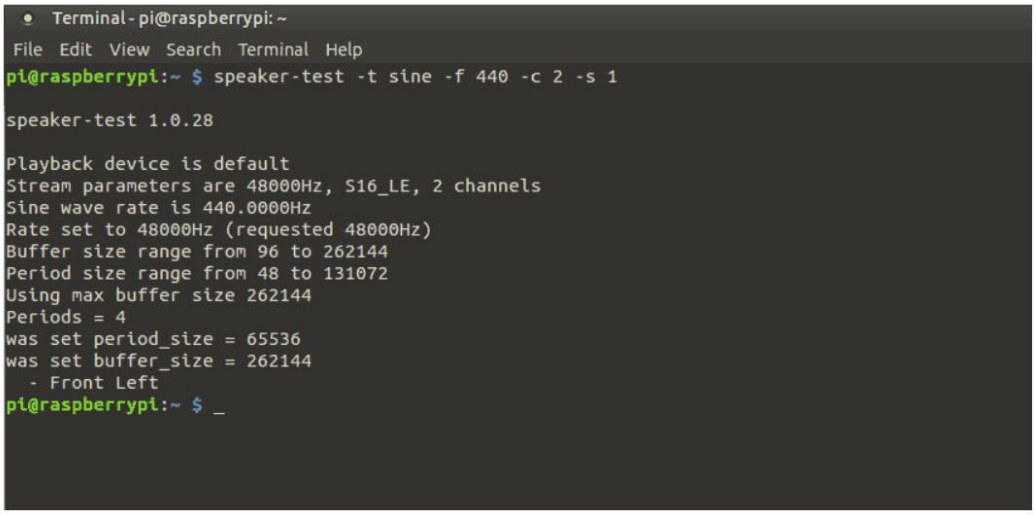

Рис. 3.16. Тестування аудіо з відображенням параметрів відтворення

Окрім того, можна використати альтернативний спосіб, запустивши утиліту mpg321 для перевірки і відтворення mp3 файлу.

3.3. Програмне забезпечення для перетворення тексту в аудіопотік

Для практичного застосування розробленої моделі перетворення тексту в аудіопотік необхідно реалізувати і налаштувати:

– транслятор, що забезпечує передачу і відтворення аудіопотоку;

– прикладне програмне забезпечення реалізації запропонованої моделі TTS;

– PubNub, що забезпечує обмін повідомленнями в реальному часі між клієнтом і відтворювачем аудіопотоку.

3.3.1. Транслятор аудіопотоку

Для забезпечення трансляції текстового повідомлення в аудіопотік розроблено скрипт за допомогою мови програмування Python, який необхідно запустити на Raspberry PI. На рис. 3.17 показано фрагмент, що реалізує підключення необхідних бібліотек та функцію ініціалізації процесу трансляції одержаного текстового повідомлення в аудіопотік.

```
from pubnub import Pubnub
import subprocess
import sys
import os

pubnub_requestchannel="audiostream_request"
pubnub_responsechannel="audiostream_response"

postive_response={"type" : "response" , "status" : "positive"}
negative_response={"type" : "response" , "status" : "negative"}
done={"type" : "response" , "status" : "done"}


FNULL = open(os.devnull,'w')

def init():
        global pubnub
        pubnub = Pubnub(publish_key="demo", subscribe_key="demo")
        pubnub.subscribe(channels=pubnub_requestchannel, callback=_callback, error=_error)
```

Рис. 3.17. Фрагмент коду ініціалізації broadcast

Для організації broadcast використовується PubNub, що є високопродуктивною комунікаційною платформою реального часу та оптимізована під максимальну пропускну здатність. Дана платформа може ефективно використовуватися при створенні чатів, систем керування IoT, складних програмно-апаратних комплексів геолокації та диспетчеризації і багатьох інших комунікаційних рішень. Схему застосування даної платформи наведено на рис. 3.18.

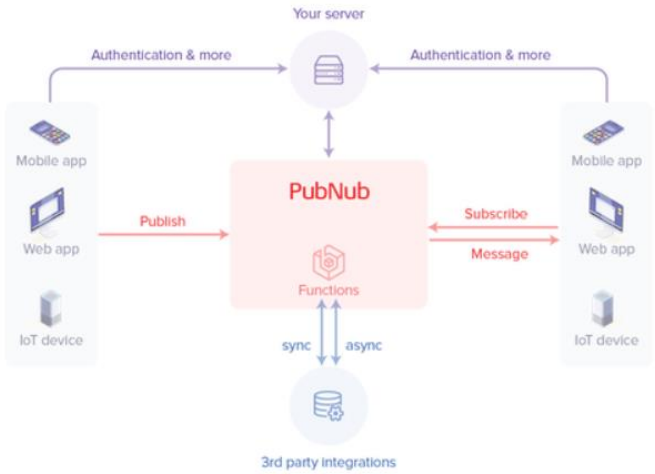

Рис. 3.18. Схема використання платформи PubNub

Функція безпосереднього відтворення аудіопотоку в залежності від результатів опрацювання текстового повідомлення реалізована як показано на рис. 3.19.

```
def _callback(message,channel):

        if message["type"]=="request" :
                print" Received message = ", message["play"]
                status=subprocess.call(["espeak","-s 120 -v en ",message["play"]], stdout=FNULL, stderr=subprocess.STDOUT)

                if status==0 :
                        pubnub.publish(pubnub_responsechannel, postive_response)
                elif status!=0 :
                        pubnub.publish(pubnub_responsechannel,negative_response)

        if message["type"]=="completed" :
                pubnub.publish(pubnub_responsechannel, done)
                sys.exit()

def _error(message):
        print("Error in pubnub")


if __name__ == '__main__':
        init()
```

Рис. 3.19. Функція відтворення аудіопотоку

3.3.2. Програмний модуль відправки текстових повідомлень

Модуль відправки текстових файлів або користувацьких повідомлень також функціонує за допомогою платформи PubNub. На рис. 3.20 проілюстровано фрагмент програмного коду для підключення необхідних бібліотек, функцій ініціалізації та зчитування файлу у форматі txt.

```
from pubnub import Pubnub
import threading
import sys

pubnub_requestchannel="audiostream_request"
pubnub_responsechannel="audiostream_response"

lock=threading.Lock()
lock.acquire()

complete_sentence={"type" : "completed" }


def init():
        global pubnub
        pubnub = Pubnub(publish_key="demo", subscribe_key="demo")
        pubnub.subscribe(channels=pubnub_responsechannel, callback=_callback, error=_error)

def importData():
        try:
                f = open ("./message.txt","r")
                read = f.read()
                f.close()
                return read

        except IOError:
                print "cannot open the file"
                sys.exit()
```

Рис. 3.20. Скрипт імпорту текстового файлу

Функція «callback» виконує перевірку коректності зчитування вмісту текстового файлу і відправки на потрібний канал його відтворення у вигляді аудіопотоку (рис. 3.21).

```
def _callback(message,channel):


        if message["status"]== "positive":
                lock.release()

        elif message["status"]== "negative":
                print "error in response"

        elif message["status"]== "done":
                sys.exit()



def _error(message):
        print("Error in pubnub")
```

Рис. 3.21. Функція «callback»

Функція для опрацювання запиту на перетворення тексту в аудіопотік представлена на рис. 3.22.

```
def process_request():
        sentence = importData().split(".")

        for data in sentence :
                print "sending message: ",data
                request={"type" : "request" , "play" : data}
                pubnub.publish(pubnub_requestchannel, request)
                lock.acquire()

        pubnub.publish(pubnub_requestchannel,complete_sentence)
```

Рис. 3.22. Функція формування запиту на відтворення аудіопотоку

Таким чином, реалізовано практичну сторону відправки і трансляції текстового повідомлення у вигляді файлу з розширенням txt за допомогою двох скриптів. Перший з них виконується на стороні Raspberry PI, а інший – запускається на комп'ютері користувача. Завершальним етапом практичної реалізації запропонованого підходу до організації комп'ютерної системи перетворення текстової інформації в аудіопотік є програмна інтеграція запропонованої у розділі 2 моделі архітектури.

3.4. Реалізація та оцінювання ефективності моделі перетворення тексту в аудіопотік

Для навчання запропонованої архітектури нейронної мережі перетворення тексту в аудіопотік використовується набір даних LJSpeech, який містить 13 100 аудіофайлів із відповідними текстовими транскриптами.

Для навчання мережі використовується 12 588 зразків, а для тестування – 512. Послідовність фонем генерується за допомогою g2p, конвертера англійських графем (орфографії) у фонеми (вимова) з відкритим кодом.

Форма хвилі перетворюється на спектрограму Мела з довжиною вікна 1024, довжиною стрибків 256 і частотою дискретизації 22050. Отримана

mel-спектрограма має 80 каналів. Показник MFA використовується для встановлення цільової тривалості фонеми.

Істинні значення висоти та енергії (амплітуди) обчислюються за допомогою STFT та WORLD вокодера відповідно. Загальна функція втрат показана у формулі 3.1:

$$\mathcal{L} = \alpha\mathcal{L}_{mel} + \beta\mathcal{L}_p + \gamma\mathcal{L}_e + \lambda\mathcal{L}_d \qquad (3.1)$$

Функція втрат на спектрограмі Мела $\mathcal{L}_{mel}$ дорівнює L_1 з $\alpha = 10$. Середньо квадратична похибка MSE використовується для функцій втрат параметра висоти звуку: $\mathcal{L}_p$, енергії $\mathcal{L}_e$ і тривалості $\mathcal{L}_d$. $\beta = 2, \gamma = 2$ і $\lambda = 1$. L = αLmel + βLp + γLe + λLd. (1)

Модель EfficientSpeech тренується протягом 5000 епох. Розмір пакету становить 128. Оптимізатором є AdamW зі швидкістю навчання 0,001, затуханням косинусної швидкості навчання та кількістю епох 50. Програмна реалізація класу для визначення моделі перетворення тексту в аудіопотік показана на рис. 3.23.

```
class EfficientSpeech(LightningModule):
    def __init__(self,
                 preprocess_config,
                 lr=1e-3,
                 weight_decay=1e-6,
                 max_epochs=5000,
                 depth=2,
                 n_blocks=2,
                 block_depth=2,
                 reduction=4,
                 head=1,
                 embed_dim=128,
                 kernel_size=3,
                 decoder_kernel_size=3,
                 expansion=1,
                 wav_path="wavs",
                 hifigan_checkpoint="hifigan/LJ_V2/generator_v2",
                 infer_device=None,
                 verbose=False):
        super(EfficientSpeech, self).__init__()

        self.save_hyperparameters()

        with open(os.path.join(preprocess_config["path"]["preprocessed_path"], "stats.json")) as f:
            stats = json.load(f)
            pitch_stats = stats["pitch"][:2]
            energy_stats = stats["energy"][:2]
```

Рис. 3.23. Фрагмент реалізації класу для побудови моделі TTS

Оцінка запропонованої моделі здійснюється не лише з точки зору якості створеного мовлення, але й компромісу щодо кількості параметрів, кількості обчислень, які вимірюються операціями з плаваючою комою (FLOPS), а також швидкості чи пропускної здатності з точки зору затримки.

Комплексний тест дає змогу отримати загальну картину продуктивності запропонованої моделі перетворення тексту в аудіопотік, як функції від пам’яті, обчислювального бюджету та часу замість того, щоб зосереджуватися лише на вибраних оптимальних показниках.

Кількість параметрів, зазвичай, використовується як проксі визначення обсягу пам'яті, необхідного моделі під час виконання. FLOPS відображає кількість операцій Fused-Multiply-Add (FMA), необхідних для завершення формування відповіді моделі.

Для змінної довжини вхідної текстової послідовності, як у TTS, FLOPS вимірюється за допомогою 128 випадкових текстових даних з тестової вибірки. Кількість операцій збільшується з довжиною введеного тексту. Затримка вимірюється кількістю секунд мовлення, згенерованого за секунду, або коефіцієнтом реального часу (RTF).

Обернений до RTF, час, необхідний для створення голосового повідомлення тривалістю 1 секунда, також можна використовувати, але це призводить до малих дробових чисел, які менш інтуїтивно зрозумілі для інтерпретації. Щоб зосередитися на швидкості запропонованої моделі, введено показник спектрограми Мела у реальному часі (mRTF). Цей показник інтерпретує це кількість секунд аудіопотоку розділений на час генерації спектрограми Мела. Метрика «fvcore» використовується для обчислення кількості параметрів і FLOPS. Для вимірювання часу використовується показник CMOS.

У табл. 3.1 показано кількість параметрів і відносний слід запропонованої моделі у порівнянні з найсучаснішими генераторами спектрограми Мела.

Таблиця 3.1

Кількість параметрів у різних моделях генератора спектрограм Мела

Модель	Абсолютний показник кількості параметрів моделі (млн)	Відносний показник кількості параметрів моделі (%) відносно EfficientSpeech
EfficientSpeech (ES)	0,27	-
FastSpeech2[1] Tacotron2	30,81 23,81	0,86% 1,12%
MixerTTS	20,06	1,33%
LightSpeech	1,80	14,78%

Модель, запропонована у роботі, є крихітною у порівнянні з іншими і включає лише 266 тис. параметрів. Така кількість параметрів вимагає дуже малої кількості операцій, як показано в табл. 3.2.

Таблиця 3.2

Об’єм обчислень в термінах GFLOPS в різних моделях генератора спектрограм Мела

Модель	Абсолютний показник, GFLOPS	Відносний показник GFLOPS відносно моделі EfficientSpeech (%)
EfficientSpeech (ES)	0,09	-
FastSpeech2	15,87	0,57%
Tacotron2	16,20	0,56%
MixerTTS	10,29	0,87%
LightSpeech	0,76	11,84%

Ефектом невеликої кількості параметрів і FLOPS є швидке створення мел-спектрограм, що досягає mRTF 953,3 на GPU V100, як показано в табл.3.3.

Таблиця 3.3

Експериментальні дані за метрикою mRTF

Модель	GPU V100	Прискорення, %	Xeon 2.2G	Прискорення, %	ARM 1.5G	Прискорення, %
EfficientSpeech	953.3	-	470.2	-	104.3	-
FastSpeech2	371.3	2.6×	64.7	7.3×	5.2	20.1×
Tacotron2	8.3	114.7×	1.2	379.4×	0.2	462.2×
MixerTTS	204.9	4.7×	55.2	8.5×	2.9	36.5×
LightSpeech	-	-	107.5	4.4×	-	-

Швидкість більш очевидна на процесорі RPi4, де запропонована модель досягає швидкості формування спектрограми Мела на рівні 104,3, що у 20,1 разів швидше порівняно з FastSpeech2.

Для Tacotron2 і MixerTTS було оцінено попередньо підготовлені версії, надані NVIDIA NeMo. При генерації аудіопотоку обидві моделі не можуть працювати з $RTF \geq 1.0$ на процесорі Raspberry PI 4. Крім того, NeMo використовував змішане навчання точності та інші оптимізації, забезпечуючи значне прискорення графічних процесорів.

У табл. 3.4 показано метрику CMOS, оцінену 15 учасниками з високим рівнем розуміння англійської на слух. Синтезовані сигнали мовлення взяті з тестового розбиття. У цьому випадку і запропонована модель, і FastSpeech2 використовували невелику версію стандартного HiFiGANv2 із 0,9 млн параметрів.

Таблиця 3.4

Значення метрики CMOS

Model	CMOS↑
FastSpeech2	0.0
EfficientSpeech	-0.14
LightSpeech	0.04

Що стосується якості аудіо, вихід запропонованої моделі зазнає лише незначного погіршення, незважаючи на невеликий розмір. Для довідки також показано опубліковану оцінку CMOS LightSpeech порівняно з FastSpeech2.

При mRTF 104,3 на RPi4 запропонована модель має значний запас для прискорення генерації аудіопотоку завдяки «легкому» вокодеру. При проведенні експерименту нейронна генеративна мережа HiFi-GAN споживає 5,0 GFLOPS, тоді як накладні витрати запропонованої моделі становлять лише 0,09.

Обчислювальна продуктивність недорогої системи RPi Zero з 256 МБ або 512 МБ оперативної пам’яті становить приблизно 0,2-0,3 GFLOPS, що дає моделі достатньо свободи дій, але не для вокодера.

Raspberry PI 3 Model B з 1 ГБ оперативної пам’яті має обчислювальну продуктивність приблизно від 3,6 до 6,2 GFLOPS, друга ж версія – має приблизно від 1,5 до 4,4. Теоретично, вокодер нижче 0,1 GFLOPS забезпечить широке впровадження нейронних моделей перетворення тексту в аудіопотік на багатьох недорогих і малопотужних пристроях.

При 266 тис. параметрів, 16-розрядному обчисленні з плаваючою точкою займана об’єм запропонованої моделі становить приблизно 532 Кб, що залишає достатньо місця в оперативній пам’яті для зберігання результатів проміжних рівнів навіть на системах з процесорами 256 МБ із малою пам’яттю.

3.5. Висновки до розділу

Основні результати даного розділу полягають в наступному:

1. Розроблено схему організації комп’ютерної системи для перетворення текстових повідомлень у аудіопотік до складу якої входять Raspberry PI, зовнішня аудіоплата та динаміки відтворення звуку.

2. Налаштовано параметри для коректного відтворення звукових сигналів Raspberry PI на системному рівні та проведено їх тестування.

3. Розроблено системний програмний додаток для забезпечення трансляції текстових повідомлень в аудіосигнал з використанням мови програмування Python, що дало змогу відправляти текстові файли з користувацького пристрою та відтворювати їх вміст у вигляді аудіосигналу на Raspberry PI.

4. Програмно реалізовано модель запропонованої архітектури нейронної мережі та проведено експериментальне порівняння з іншими моделями, що дало можливість довести доцільність реалізації системи, оскільки у запропонованій моделі на порядок менше параметрів та існує можливість ефективно функціонувати на пристроях з обмеженими ресурсами, зокрема Raspberry PI, починаючи з версії 2.

РОЗДІЛ 4
ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1. Охорона праці

У кваліфікаційній роботі магістра досліджено методи та інструменти побудови комп’ютерних систем аналізу і перетворення текстових повідомлень в аудіопотік. Оскільки, проектування засобу трансформації тексту у звук виконується за допомогою ПК та IoT пристроїв, то важливим аспектом роботи користувача є його безпека. У зв’язку з цим, необхідно проаналізувати і врахувати вимоги і норми охорони праці, а також правила техніки безпеки при використанні електронно-обчислювальних засобів і периферійних пристроїв. На сьогодні основним нормативним документом, який визначає і регламентує норми і правила експлуатації електронно-обчислювальної техніки є НПАОП 0.00-7.15-18 «Вимоги щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями» [22]. Правила встановлюють вимоги безпеки до обладнання робочих місць операторів ЕОМ.

НПАОП 0.00-7.15-18 є обов'язковим для виконання роботодавцями, операторами електронно-обчислювальних машин, операторами комп'ютерного набору, операторами комп'ютерної верстки та працівників інших професій, які у своїй роботі застосовують ЕОМ з ВДТ і ПП [22].

Згідно НПАОП 0.00-7.15-18 електронно-обчислювальні засоби повинні відповідати вимогам чинних в Україні стандартів і пройти державну санітарно-епідеміологічну експертизу у Порядку проведення державної санітарно-епідеміологічної експертизи.

З метою забезпечення електробезпеки користувачів ПК при проектуванні комп’ютерного засобу перетворення тексту в аудіопотік необхідно, щоб комп’ютери і периферійні пристрої відповідали І-му класу захисту, або були заземленими відповідно до вимог НПАОП 40.1-1.32-01.

Неприпустимим є використання клем функціонального заземлення для підключення захисного заземлення [23].

При організації робочих місць користувачів комп'ютерних систем необхідно забезпечити дотримання вимог щодо їх розташування, зокрема відстань робочого місця від стіни повинна складати 1 м, а відстань між робочими місцями повинна становити 1,7 м. Площа, яка виділяється на одне робоче місце, обладнане ПК становить – ≥6.0 м^2, а об'єм – ≥20 м^3 [22].

При виборі кімнат для розміщення робочих місць ПК враховано ступінь відбиття світла на екранах дисплеїв, яке проходить через вікна і яке може викликати значне осліплення в тих, хто сидить перед ними, особливо влітку та в сонячні дні. Тому, ПК і оргтехніка розміщені біля стін, які не знаходяться біля вікон або навпроти них [23].

Оскільки, при незадовільному освітленні знижується продуктивність праці користувачів ПК, і можливі негативні впливи на здоров'я такі, як короткозорість, швидка втомленість, тому всі приміщення, які облаштовані робочими місцями з ПК, мають природне і штучне освітлення. Не допускається розташування робочих місць з ПК в підвальних приміщеннях [23].

Робочі місця з ПК при виконанні творчої роботи, яка потребує значної розумової концентрації, зокрема при реалізації алгоритмів машинного навчання для перетворення тексту в аудіопотік, ізольовано одне від одного перегородкою висотою 1,6 м [23]. Поверхня підлоги у приміщеннях повинна бути оздоблена керамічною плиткою і бути рівною та зручною для очищення та вологого прибирання.

Штучне освітлення у приміщеннях повинно бути виконано у вигляді комбінованої системи освітлення з використанням люмінесцентних джерел світла у світильниках загального освітлення, які розташовувати над робочими поверхнями у рівномірно-прямокутному порядку. Штучне освітлення забезпечує на робочих місцях з ПК освітленість 300 – 500 Лк [23].

Для запобігання засвітленню екранів ПК прямими світловими потоками лінії світильників розташовані з достатнім бічним зміщенням відносно рядів робочих місць, а також паралельно до світлових отворів. При цьому кожне вікно повинно мати світлорозсіюючі штори з коефіцієнтом відбивання 0,7 [22].

У приміщенні також необхідно забезпечити і природнє освітлення, при цьому на кожному вікні закріплені жалюзі з вертикальними ламелями, що регулюються для зменшення прямого попадання сонячного світла на екран комп'ютерів.

З метою запобігання нещасним випадкам та організації охорони праці на виробництві розробляються інструкції з охорони праці і техніки безпеки при використанні комп'ютерної техніки. Дія інструкції поширюється на всі структурні підрозділи установи [22].

До роботи на ПК допускаються особи, які пройшли спеціальне навчання, медичне обстеження, вступний інструктаж з охорони праці, інструктаж на робочому місці та інструктаж з пожежної безпеки [22].

З ергономічної точки зору, при розташуванні елементів робочого місця враховано наступні фактори [22]:

– простір для розміщення користувача;
– можливість огляду елементів робочого місця;
– можливість огляду простору за межами робочого місця;
– можливість робити записи, розміщення документації і матеріалів, які використовує користувач.

При розробці методу та інструменту перетворення текстових даних в аудіопотік проведено аналітичний огляд чинних вимог з охорони праці і техніки безпеки та відповідно враховано їх, що дозволило забезпечити зручні умови для ефективної роботи користувачів спроектованої комп'ютерної системи.

4.2. Засоби захисту персоналу від уражень радіації

4.2 Планування заходів цивільного захисту на об’єкті у випадку надзвичайних ситуацій

Враховуючи необхідність проведення державних заходів захисту населення від надзвичайних ситуацій прийнято Закон "Про цивільну оборону України". Згідно із Законом кожен громадянин має право на захист свого життя і здоров'я від наслідків аварій, катастроф, пожеж, стихійного лиха та має право на надання гарантій забезпечення реалізації цього права від Кабінету Міністрів України, міністерств та інших центральних органів місцевого самоврядування, керівництва підприємств, установ і організації незалежно від форм власності й підпорядкування. Як гарант цього права держава створила систему цивільної оборони.

З метою забезпечення стійкості роботи важливих виробничих підприємств (об'єктів) в надзвичайних ситуаціях мирного і воєнного часу завчасно проводиться комплекс організаційних і інженерно-технічних заходів цивільної оборони, спрямованих на забезпечення захисту населення і зменшення руйнувань, на підвищення стійкості роботи об'єкту, на утворення необхідних умов для проведення РіІНР [24].

Основними вимогами до проектування й побудови промислових підприємств (об'єктів є забезпечення стійкості інженерно-технічного комплексу об'єкту.

Згідно [24], будівлі і споруди на об'єкті необхідно розміщувати розосереджено, при цьому відстань між будівлями повинна забезпечувати протипожежні розриви.

Пожежні розриви між будівлями повинні бути проектовані таким чином, щоб виключити можливість перенесення вогню з однієї будівлі на іншу В загальному випадку, ширина протипожежного розриву *Lp, м*, визначається за формулою: $Lp=H1 + H2 +15$ *м*, де *H1* і *H2* – висота сусідніх

будинків. Будівлі адміністративно-господарського і обслуговуючого призначення повинні розміщуватись окремо від основних цехів.

Виходячи з того, що важливі виробничі споруди будують заглибленими або пониженої висоти, і вони, зазвичай, прямокутної форми, то виробничі цехи повинні бути спроектовані у відповідності до цих вимог. Це дозволяє зменшити парусність будівлі і збільшити її опір при ударній хвилі будь-якого вибуху. Висока стійкість до дії ударної хвилі властива залізобетонній будівлі з металевими каркасами в бетонній опалубці.

Для підвищення стійкості до пожеж необхідно застосовувати вогнестійкі конструкції та проведення вогнезахисної обробки горючих елементів будівлі. У цехах, які збудовано з каменю перекриття виготовляється з бетонних плит.

У технічних заходах пожежного захисту повинно передбачатися 4 виходи, ширина коридорів у приміщеннях повинна складати 2,6 м. Усі будівлі підприємств повинні використовувати систему протидимового захисту, пожежного зв’язку і сигналізації, а будівельні роботи повинні бути виконані з вогнетривких матеріалів. Усі приміщення повинні бути обладнані засобами пожежогасіння.

Організаційні заходи проектування на підприємствах реалізуються шляхом навчання працюючих правилам пожежної безпеки, розробки інструкцій про правила роботи з пожежонебезпечими матеріалами та про дії персоналу під час пожежі.

У ряді випадків при проектуванні і будівництві промислових будівель і споруд повинна бути передбачена можливість герметизації приміщень від проникнення радіоактивного пороху. Це особливо важливо для підприємств харчової промисловості і продовольчих складів.

Деякі унікальні види технологічного устаткування потрібно розміщувати в більш міцних спорудах (підвалах, підземних спорудах) або будівлях з легких незгоряючих конструкцій павільйонного типу, під навісами або відкрито. Це обумовлюється тим, що в багатьох випадках

устаткування може витримати набагато більший надлишковий тиск ударної хвилі, ніж будівля, в якій воно знаходиться. При зруйнуванні будівлі внаслідок падіння конструкцій розміщене в них устаткування буде виходити з ладу.

Дороги на території підприємств повинні мати тверде покриття і забезпечувати зручний шлях руху між сусідніми будівлями; кількість в'їздів на територію даного підприємства об'єкту – два у різних напрямках. Системи побутової і виробничої каналізації повинні мати два випуски у міську каналізаційну мережу.

Електрозабезпечення є основою будь-якого виробництва. Порушення нормальної подачі електроенергії на об'єкт або окремі ділянки виробництва може призвести до повної зупинки роботи об'єкту.

Для надійного електрозабезпечення підприємстві в умовах надзвичайних ситуації при її проектуванні і будівництві повинні бути враховані основні вимоги, які випливають із завдань цивільної оборони. Електрозабезпечення підприємств повинно здійснюватись від енергосистем міста, до складу яких входять електростанції, що працюють на різних видах палива. При електропостачанні об'єкту від одного джерела передбачено два вводи з різних напрямів. Електроенергію на ділянках виробництва подають по належних електрокабелях, прокладених в землі на глибині 1 м.

Крім цього, на території підприємств необхідно передбачити створення автономних резервних джерел електропостачання у вигляді автономних генераторів на дизельному паливі.

Для підвищення стійкості постачання об'єктів водою необхідно, щоб система водопостачання здійснювалась не менше, ніж від двох незалежних джерел, одне з яких бажано влаштовувати підземним.

Стійкість мережі водопостачання підвищується при заглибленні в ґрунт всіх ліній водопроводу і розташування належних гідрантів і відключаючих пристроїв на території, яка не може бути заваленою, а також

пристроїв перемичок, які дозволяють відключити пошкоджені ланки і споруди.

На багатьох виробничих об'єктах газ використовується як паливо, а на хімічних підприємствах - і як вихідна сировина. При порушенні мережі, газ може стати причиною вибуху, пожежі. Для більш надійного постачання газ повинен подаватися в місто і на промислові об'єкти по двох незалежних газопроводах. Газорозподільчі станції повинні бути розташованими за межами міста. Газова мережа закільцьовується і прокладається під землею на глибині 0,6-1,7 м. На газовій мережі у визначених місцях встановлюються автоматичні відключаючі пристрої, які спрацьовують від надлишкового тиску ударної хвилі.

Крім того, на газопроводах встановлюють відключаючу апаратуру з дистанційним управлінням і крани, які автоматично перекривають подачу газу при розриві труб, що дозволяє відключити газові мережі певних ділянок і районів міста.

4.2 Організація робочого місця користувача відеодисплейним терміналом

Робота відеотерміналів включає різні завдання, які об'єднуються такими загальними чинниками, як те, що робота проводиться в сидячому положенні і вимагає уважного, неперервного і іноді тривалого спостереження.

Виділяють три групи основних завдань, які розв'язуються на відеотерміналах:

– контроль і спостереження;
– діалог;
– збір інформації.

Ці завдання розрізняються по тривалості використання дисплея і по ступеню уваги, якого вони вимагають. Важливим питанням є режим праці і

відпочинку при роботі з відеотерміналами. Виділяють 7 умов для того, щоб діяльність на робочому місці, оснащеному дисплеєм, здійснювалася без скарг і без втоми [23].

Правильне облаштування робочого столу:

– при фіксованій висоті – оптимальна висота - 720мм;

– повинен забезпечуватися необхідний простір для рук по висоті, ширині і глибині;

– в області сидіння не повинно бути шухляд.

Правильне встановлення робочого стільця:

– висота повинна регулюватися;

– конструкція повинна бути такою, що обертається;

– правильна висота сидіння: площа сидіння на 30мм нижче, ніж підколінна западина.

Правильне розташування приладів: необхідно так установити яскравість знаків і яскравість фону дисплея, щоб не було великої відмінності в порівнянні з яскравістю навколишнього оточення, але щоб знаки чітко пізнавалися на відстані читання. Не допускати [22]:

– дуже велику яскравість (викликає мерехтіння);

– дуже слабку яскравість (сильне навантаження на очі);

– дуже чорну фонову яскравість дисплея (сильне навантаження на очі).

Правильне виконання робіт:

– положення тулуба пряме, ненапружене;

– положення голови пряме, вільне, зручне;

– положення рук - зігнуті трохи більше, ніж під прямим кутом;

– положення ніг - зігнуті трохи більше, ніж під прямим кутом;

– правильна відстань для зору, клавіатура і дисплей –приблизно на однаковій відстані для зору: при постійній роботі - близько 500мм, при випадковій роботі - до 700мм.

Правильне освітлення:

– освітлення по можливості із сторони, зліва;

– по можливості - рівномірне освітлення всього робочого простору;

– прилади по можливості встановлювати в місцях, віддалених від вікон;

– вибирати непряме освітлення приміщення або вкривати корпуси світильників;

– світло, що поступає через вікна, пом'якшувати за допомогою штор;

– організувати робоче місце, щоб напрям погляду йшов по можливості паралельно фронту вікон.

Правильне застосування допоміжних засобів: підлокітники використовувати, якщо клавіатура вища 15мм [23].

Правильний метод роботи:

– передбачати по можливості зміну завдань і навантажень;

– дотримувати перерви в роботі: 5 хвилин через 1 годину роботи біля дисплея або 10 хвилин після 2-х годин роботи біля дисплея.

Ергономічна організація робочого місця користувача ЕОМ повинна враховувати як специфіку діяльності, що виконується, так забезпечувати комфортні умови перебування людини.

Тому до основних ергономічних завдань щодо організації робочого місця слід віднести [23]:

– забезпечення просторових параметрів робочого місця, які відповідають антропометричним характеристикам користувача;

– раціональне розташування елементів робочого місця відносно користувача на підставі поглибленого кількісного та якісного аналізу діяльності, яка виконується;

– оптимізацію умов робочого середовища.

В ході організації робочих місць на кожну ЕОМ повинна бути виділена площа, яка складає не менш, ніж 6 $м^2$, та об'єм, який становить не менш, ніж 24 $м^3$. Причому, зона, де розташовується робочий стіл, сервер або

робоча станція, принтер, екран для графопроектора, повинна займати відповідно 6 - 8 м^2. Висота приміщення повинна бути не менш, ніж 4 м [23].

Робоче місце користувача ПК повинно бути обладнане одномісним столом та напівм'яким стільцем, висоту сидіння яких можна змінювати. Довжина стола повинна бути не менше 70 см, ширина – забезпечувати місце перед клавіатурою (не менше, ніж 40 см) для розташування зошита або іншого приладдя. Поверхня стола повинна мати кут нахилу у межах 12-150, лише іноді припустимою є її розташування у горизонтальній площині.

Слід забезпечити відповідність висоти краю стола, що повернений до користувача, і стільця над підлогою росту та антропометричним особливостям організму користувачів.

Глибина простору для ніг під столом повинна бути не менше 45 см, а у випадку застосування високого стола та низького стільця і, отже, відсутності відповідності росту користувача конструктивним елементам робочого місця, слід використовувати підставку для ніг, ширина якої становить – 35 см, довжина – 40 см, кут нахилу опорної поверхні – 15^o.

Столи з ЕОМ можуть бути розміщені без розривів між ними, а при незначній кількості робочих столів з відеотерміналами перевагу слід віддавати розташуванню їх біля внутрішньої стіни.

Робота з комп'ютерною технікою вимагає обов'язкового дотримування правильної посадки. Користувач ЕОМ повинен сидіти прямо, з невеликим нахилом (до 5 – 7°) голови вперед, не сутулитися, спираючись нижніми краями лопаток на спинку стільця. Передпліччя повинні спиратися на поверхню стола, забезпечуючи зниження статичного напруження м'язів плечового поясу і рук, кути, що утворюються передпліччям і плечем, а також гомілкою і стегном, – складати не менш, ніж 90°.

Рівень очей повинен припадати на центр екрана або на точку, яка розташована між верхньою та середньою третинами екрану, причому, лінія погляду повинна бути перпендикулярною до площини екрана, а її відхилення у вертикальній площині – знаходитися у межах ±5–10^0. Оптимальний огляд

у горизонтальній площині від центральної осі екрана повинен бути у межах ±15–30°. Лише під час спостереження за інформацією, яка розміщена у найвіддаленіших ділянках екрану, кут огляду може становити 40–45°.

Оптимальна відстань від очей до площини екрана монітора повинна складати 60–70 см, припустима – не менше 50 см. Розглядати інформацію на екрані з відстані менш, ніж 50 см не рекомендується.

Висновки.

Провівши аналіз вимог до планування заходів цивільного захисту на підприємствах у випадку надзвичайних ситуацій, стійкості об’єктів народного господарства до надзвичайних ситуацій різної природи, можна зробити висновок про необхідність дотримання чинних норм безпеки життєдіяльності та цивільної оборони. Вимоги, норми проектування, інженерно-технічні заходи цивільної оборони повинні дотримуватись в повному обсязі, що сприяє нормальному функціонуванню підприємства і надає захист працівникам не тільки в надзвичайних умовах мирного і воєнного часу, але й покращує умови праці людей. Крім того, важливим аспектом при роботі з відеодисплейними терміналами є правильна організація робочого, що забезпечує зниження втомлюваності оператора і дозволяє уникнути проблем зі здоров’ям.

ВИСНОВКИ

Основні наукові та практичні результати полягають в наступному.

1. Проведено аналіз підходів і технологій синтезу голосових повідомлень і встановлено, що автоматизація цього процесу потребує вдосконалення процесів аналізу текстової інформації, побудови більш ефективних акустичних моделей та механізмів відтворення звуку.

2. На основі аналізу таксономії процесів перетворення текстових повідомлень в аудіопотік визначено потенційні способи розвитку існуючих нейромережевих моделей, зокрема, в контексті застосування методів машинного навчання для підвищення якості попереднього опрацювання тексту, перетворення графем у фонеми, а також забезпечення можливості їх прогнозування на основі попередньо навчених нейронних моделей.

3. Проаналізовано методи токенізації тексту, визначено основні переваги і недоліки різних видів токенізації, що дало можливість обґрунтувати доцільність застосування архітектури нейронних мереж на основі трансформерів для забезпечення швидкості та ефективності майже в реальному часі.

4. Визначено основні характеристики аудіосигналів, які можна ефективно використати при перетворенні тексту у звук, що дало змогу врахувати, зокрема, спектрограми Мела при побудові компонентів архітектури нейронної мережі перетворення тексту в аудіопотік.

5. Досліджено алгоритми побудови вокодерів, які використовуються для перетворення вихідної спектрограми акустичної моделі у цільову звукову форму, що дало змогу обґрунтувати застосування GAN вокодерів при проектуванні архітектури нейронної мережі перетворення текстової інформації в аудіопотік і розпаралелити його.

6. Запропоновано архітектуру нейронної мережі до складу якої входить енкодер на базі трансформерів, які забезпечують зменшення розмірності вхідної матриці фонем у 4 рази та добувають фонетичні

властивості, а також декодера, який сформований з блоків визначення акустичних властивостей, зокрема енергії, тривалості і висоти звуку, та блоків декодування властивостей аудіосигналу на основі якого відбувається генерація спектрограми Мела, що дало змогу забезпечити високу продуктивність і якість формування голосового повідомлення у порівнянні з іншими моделями.

7. Розроблено схему організації комп'ютерної системи для перетворення текстових повідомлень у аудіопотік до складу якої входять Raspberry PI, зовнішня аудіоплата та динаміки відтворення звуку.

8. Налаштовано параметри для коректного відтворення звукових сигналів Raspberry PI на системному рівні та проведено їх тестування.

9. Розроблено системний програмний додаток для забезпечення трансляції текстових повідомлень в аудіосигнал з використанням мови програмування Python, що дало змогу відправляти текстові файли з користувацького пристрою та відтворювати їх вміст у вигляді аудіосигналу на Raspberry PI.

10. Програмно реалізовано модель запропонованої архітектури нейронної мережі та проведено експериментальне порівняння з іншими моделями, що дало можливість довести доцільність реалізації системи, оскільки у запропонованій моделі на порядок менше параметрів та існує можливість ефективно функціонувати на пристроях з обмеженими ресурсами, зокрема Raspberry PI, починаючи з версії 2.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Microsoft Azure: Text to speech. URL: https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/ (дата звернення: 05.09.2023).

2. Kovtun V., Kovtun O. System of methods of automated cognitive linguistic analysis of speech signals with noise, Multimedia Tools and Applications. Springer Science and Business Media LLC. 2022. URL: https://doi.org/10.1007/s11042-022-13249-5 (дата звернення: 08.09.2023).

3. Kovtun V., Kovtun O., Semenov A. Entropy-Argumentative Concept of Computational Phonetic Analysis of Speech Taking into Account Dialect and Individuality of Phonation, Entropy. Vol. 24, no. 7. 2022. URL: https://doi.org/10.3390/e24071006 (дата звернення: 06.09.2023).

4. Krak Y., Barmak O., Mazurets O. The practice implementation of the information technology for automated definition of semantic terms sets in the content of educational materials. In: CEUR Workshop Proceedings 2139. 2018. pp. 245-254.

5. Kryvonos I.G., Krak Iu.V., Barmak O.V., Bagrii R.O. New Tools of Alternative Communication for Persons with Verbal Communication Disorders. Cybern. Syst. Anal. 52(5). 2016. PP. 655–673.

6. Rashkevych Y., Peleshko D., Pelekh I., Izonin I. Speech signal marking on the base of local magnitude and invariant segmentation. Mathematical Modeling and Computing. 2014. 1(2), pp. 234–244.

7. Google Cloud: Text to speech. URL: https://cloud.google.com/text-to-speech (дата звернення: 06.09.2023).

8. Cerence/Nuance TTS Ukrainian. URL: https://nextup.com/cerence/ (дата звернення: 08.09.2023).

9. Mel-spectrogram. URL: https://en.wikipedia.org/wiki/Mel_scale (дата звернення: 08.09.2023).

10. Griffin-Lim Algorithm. URL: https://paperswithcode.com/method/griffin-lim-algorithm (дата звернення: 10.09.2023).

11. Луцків А.М., Макогон С.В. Нейромережеві підходи до перетворення текстових повідомлень в аудіопотік. Матеріали XII міжнародної науково-практичної конференції молодих учених та студентів «Актуальні задачі сучасних технологій» (6-7 грудня 2023 року). Тернопіль: ТНТУ. 2022. С. 438.

12. Луцків А.М., Макогон С.В. Типи архітектур нейронних мереж для перетворення текстових повідомлень у звуковий потік. Матеріали XI науково-технічної конференції Тернопільського національного технічного університету імені Івана Пулюя «Інформаційні моделі, системи та технології» (13-14 грудня 2023 року). Тернопіль: ТНТУ. 2022. С. 164.

13. J. Shen, et al. TTS Synthesis by Conditioning Wavelet on Mel Spectrogram Predictions. URL: https://arxiv.org/pdf/1712.05884.pdf (дата звернення: 26.09.2023).

14. Y. Ren, C. Hu, X. Tan, T. Qin. FastSpeech2: Fast and High-quality End-to-end Text to Speech. URL: https://arxiv.org/pdf/2006.04558.pdf (дата звернення: 05.09.2023).

15. Y. Ren, et al., Fastspeech: Fast robust and controllable text to speech, Advances in Neural

16. Information Processing Systems. URL: https://proceedings.neurips.cc/paper/2019/file/f63f65b503e22cb970527f23c9ad7db1-Paper.pdf (дата звернення: 11.09.2023).

17. Паламар М.І., Стрембіцький М.О., Паламар А.М. Проектування комп’ютеризованих вимірювальних систем і комплексів. Навчальний посібник. Тернопіль: ТНТУ. 2019. 150 с.

18. Погребенник В. Д., Клим Г. І., Бордун І. М., Пташник В. В., Паламар А. М. Системи оперативного контролю інтегральних параметрів водного середовища. Т. 2. Елементи комп’ютерних систем оперативного

контролю: колективна монографія. Житомир: Видавничий дім «Бук-Друк», 2021. 180 с.

19. K. Cho, at al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. URL: https://arxiv.org/abs/1406.1078 (дата звернення: 09.09.2023).

20. Bahdanau, K. Cho, Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. URL: https://arxiv.org/abs/1409.0473 (дата звернення: 12.09.2023).

21. Жидецький В.Ц. Охорона праці користувачів комп´ютерів. Львів: Афіша, 2011. 176 с.

22. Желібо Е.Н. Безпека життєдіяльності: Навчальний посібник/ За редакцією Е.П. Желібо, В.М. Львів: «Новий світ - 2000», 2011. 320с.

23. Стадник І.Я., Зварич Н.М. Оцінка хімічної обстановки при аваріях на хімічно небезпечних об’єктах викидом (виливом) небезпечних хімічних речовин та застосуванні хімічної зброї. ТНТУ. 2020. 36 С.

Додаток А

Текст наукових публікацій кваліфікаційної роботи магістра

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Тернопільський національний технічний університет імені Івана Пулюя (Україна)
Університет імені П'єра і Марії Кюрі (Франція)
Маріборський університет (Словенія)
Технічний університет у Кошице (Словаччина)
Вільнюський технічний університет ім. Гедимінаса (Литва)
Міжнародний університет цивільної авіації (Марокко)
Наукове товариство ім. Т.Шевченка

АКТУАЛЬНІ ЗАДАЧІ СУЧАСНИХ ТЕХНОЛОГІЙ

Збірник
тез доповідей

XII Міжнародної науково-практичної конференції молодих учених та студентів
6-7 грудня 2023 року

УКРАЇНА
ТЕРНОПІЛЬ – 2023

Матеріали XII Міжнародної науково-практичної конференції молодих учених та студентів
«АКТУАЛЬНІ ЗАДАЧІ СУЧАСНИХ ТЕХНОЛОГІЙ» – Тернопіль, 6-7 грудня 2023 року

УДК 004.048

А. М. Луцків канд. техн. наук, доцент, С. В. Макогон
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

НЕЙРОМЕРЕЖЕВІ ПІДХОДИ ДО ПЕРЕТВОРЕННЯ ТЕКСТОВИХ ПОВІДОМЛЕНЬ В АУДІОПОТІК

A. M. Lutskiv PhD., Assoc. Prof., S. V. Makohon

NEURAL NETWORK APPROACHES TO CONVERTING TEXT MESSAGES INTO AUDIO STREAMS

Застосування нейромережевого підходу для перетворення тексту в аудіо можна класифікувати в основному з точки зору основних компонентів: аналіз тексту, акустичні моделі, вокодери та повністю наскрізні моделі, як показано на рис 1. Вони формують таксономію та основні поняття при побудові алгоритмів щодо практичної реалізації методів перетворення тексту у голосові повідомлення.

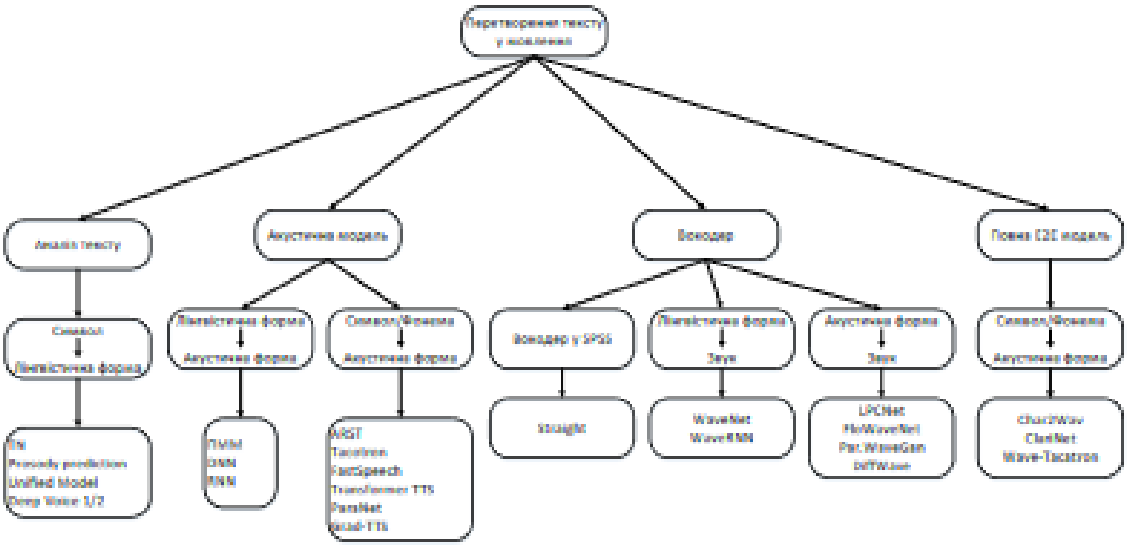

Рисунок 1. Класифікація компонентів та способів їх реалізації із застосуванням нейронних мереж при перетворенні тексту в аудіо

Така таксономія узгоджується з потоком перетворення даних із тексту в сигнал:

– аналіз тексту перетворює символ у фонему або інші мовні властивості;

– акустичні моделі формують акустичні ознаки або з мовних властивостей, або з символів/фонем;

– вокодери генерують хвилю з мовних або акустичних властивостей;

– повністю наскрізні моделі безпосередньо перетворюють символи/фонеми у хвилю.

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ
УНІВЕРСИТЕТ ІМЕНІ ІВАНА ПУЛЮЯ

М А Т Е Р І А Л И

XI НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

«ІНФОРМАЦІЙНІ МОДЕЛІ, СИСТЕМИ ТА ТЕХНОЛОГІЇ»

13-14 грудня 2023 року

ТЕРНОПІЛЬ
2023

УДК 004.048
Андрій Луцків канд. техн. наук, доцент, Сергій Макогон
Тернопільський національний технічний університет імені Івана Пулюя

ТИПИ АРХІТЕКТУР НЕЙРОННИХ МЕРЕЖ ДЛЯ ПЕРЕТВОРЕННЯ ТЕКСТОВИХ ПОВІДОМЛЕНЬ У ЗВУКОВИЙ ПОТІК

Andriy Lutskiv PhD., Assoc. Prof., Serhii Makohon

TYPES OF NEURAL NETWORK ARCHITECTURES FOR TEXT TO SPEECH

Сьогодні існують три основні типи архітектури, які використовують акустичні алгоритми. До них належать:

- алгоритми на основі рекурентних нейронних мереж;
- алгоритми на основі згорткових нейронних мереж;
- алгоритми на основі трансформерів.

Рекурентна нейронна мережа може використовуватися для представлення такої структури акустичної моделі та таких алгоритмів, як неавторегресійний Tacotron 2, структуру якого показано на рис. 1.

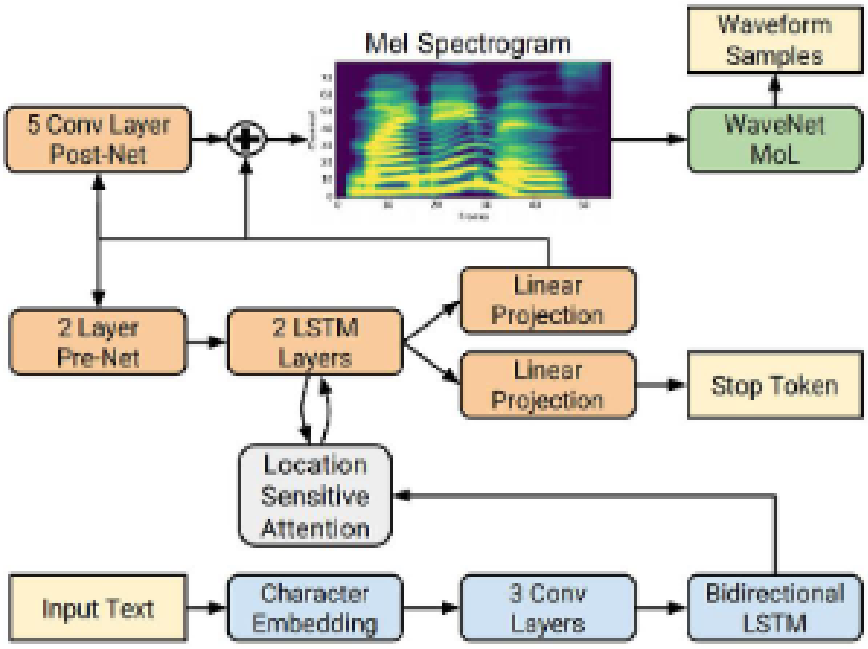

Рис. 1. Архітектура Tacotron 2

По факту, RNN мережа виконує прогнозування ознак «sequence-to-sequence», яка в свою чергу забезпечує прогнозування послідовності кадрів Mel-спектрограми з послідовності вхідних символів.

Модифікована версія WaveNet генерує зразки хвилі у часовій області, обумовлені прогнозованими фреймами Mel-спектрограми. Архітектура TacoTron 2 стала величезним кроком вперед у покращенні якості голосу порівняно з іншими методами, такими як конкатенативний, параметричний і авторегресійний TacoTron 1.