НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Радіотехнічний факультет

Кафедра радіоінженерії

«На правах рукопису»

УДК _____

До захисту допущено:

Завідувач кафедри

_____ Сергій Мартинюк

«___»_____2024р.

Магістерська дисертація

на здобуття ступеня магістра

за освітньо-науковою програмою «Радіоелектронна інженерія» зі спеціальності 172 «Телекомунікації та радіотехніка»

на тему: «Система захищеного голосового зв'язку»

Виконав:

студент ІІ курсу, групи РІ-21мн

Томчук Іван Вікторович _____

Науковий керівник:

Ст.викл.

Павлов Олег Ігорович _____

Рецензент:

Приходько Ірина Олександрівна _____

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

Студент _____

Київ – 2024 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Радіотехнічний факультет
Кафедра радіоінженерії

Рівень вищої освіти – другий (магістерський)
Спеціальність – 172 «Телекомунікації та радіотехніка»
Освітньо-наукова програма «Радіоелектронна інженерія»

ЗАТВЕРДЖУЮ
В.о. зав. кафедри
_____ Сергій МАРТИНЮК
«___»_____20__ р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Томчук Івану Вікторовичу

1. Тема дисертації «Система захищеного голосового зв'язку», науковий керівник дисертації Павлов Олег Ігорович, ст.викл., затверджені наказом по університету від «28»___03__ 2024 р. №1483-с

2. Термін подання студентом дисертації _____

3. Об'єкт дослідження *Процеси перетворення сигналів в голосовому каналі системи GSM; Процеси параметричного перетворення мовленнєвих сигналів в вокодерних системах та системах speech-to-text та text-to-speech; Методи модуляції даних для їх передачі через голосові канали системи GSM*

4. Предмет дослідження *Системи speech-to-text та text-to-speech, можливість їх наскрізного потокового застосування в реального часі з мінімальними наскрізними затримками для забезпечення роботи з різними мовами, можливості криптографічного p2p шифрування псевдотексту, можливості перетворення псевдотексту на speech-like сигнали, можливості передачі speech-like сигналів через голосовий канал системи GSM при використанні кодеків AMR-4750 з мінімальною швидкістю передачі параметричних даних*

РЕФЕРАТ

Магістерська дисертація: 96 с., 26 рис., 2 додатки, 13 таблиць.

Тема дисертації: «Система захищеного голосового зв'язку»

Об'єкт дослідження — Процеси перетворення сигналів в голосовому каналі системи GSM; процеси параметричного перетворення мовленнєвих сигналів в вокодерних системах та системах speech-to-text та text-to-speech; методи модуляції даних для їх передачі через голосові канали системи GSM.

Мета дослідження — Отримання інформації про можливість застосування систем STT та TSS для наскрізного потокового кодування багатомовних мовленнєвих сигналів в реальному часі, можливість інтеграції таких систем в систему передачі шифрованого мовлення через голосові низькошвидкісні канали GSM, можливість реалізації таких систем на базі портативних мікроелектронних пристроїв, оцінювання якісних і кількісних характеристик роботи таких систем в означених умовах.

Отримані результати — Концепція побудови системи передачі шифрованого мовлення через голосові канали GSM з використанням систем STT та TSS на її вході і виході. Адаптовані версії ПЗ для реалізації систем на базі портативних мікроелектронних пристроїв. Якісні і кількісні характеристики роботи систем STT та TSS при наскрізному потоковому кодуванні багатомовних мовленнєвих сигналів в реальному часі.

Ключові слова: захищений голосвий зв'язок, низькошвидкісне кодування мовлення, системи speech-to-text та text-to-speech, голосовий канал GSM, speech-like модем, затримка синтезу-розпізнавання мовлення, потокове розпізнавання та синтез мовлення, p2p захист голосового каналу AMR-4750.

ABSTRACT

Master's dissertation: 96 p., 26 figures, 2 appendices, 13 tables.

Dissertation topic: « Secure voice communication system ».

Object of study: Signal conversion processes in the voice channel of the GSM system; processes of parametric transformation of speech signals in vocoder systems and speech-to-text and text-to-speech systems; methods of data modulation for their transmission through voice channels of the GSM system.

The aim of the study: obtaining information about the possibility of using STT and TSS systems for end-to-end streaming coding of multilingual speech signals in real time, the possibility of integrating such systems into the system of transmission of encrypted speech via low-speed GSM voice channels, the possibility of implementing such systems based on portable microelectronic devices, evaluating qualitative and quantitative characteristics operation of such systems under the specified conditions.

Research goal: the concept of building an encrypted broadcast transmission system through GSM voice channels using STT and TSS systems at its input and output. Adapted software versions for implementation of systems based on portable microelectronic devices. Qualitative and quantitative performance characteristics of STT and TSS systems during end-to-end stream coding of multilingual speech signals in real time.

Keywords: secure voice communication, low-speed speech coding, speech-to-text and text-to-speech systems, GSM voice channel, speech-like modem, delayed synthesis-speech recognition, stream recognition and speech synthesis, p2p protection of the voice channel AMR- 4750.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

PSOLA – Pitch Synchronous Overlap Add (синхронне перекриття додавання висоти тону);

SPSS – Statistical Parametric Speech Synthesis (статистичний параметричний синтез мови);

RNN – рекурентні нейронні мережі;

CNN – згорткові нейронні мережі;

LSTM – Long Short-Term Memory (довготривала короткочасна пам'ять);

GRU – Gated Recurrent Units (рекурентні блоки з вентилями);

FPGA – Field-Programmable Gate Array (польова програмована вентильна матриця);

ASIC – Application-Specific Integrated Circuit (інтегральна схема спеціального призначення);

CRM – Customer Relationship Management (управління взаємовідносинами з клієнтами);

MOS – Mean Opinion Score (середній оцінний бал);

PESQ – Perceptual Evaluation of Speech Quality (перцептивна оцінка якості мови);

VRR – Viseme Recognition Rate (рівень розпізнавання візем);

MUSHRA – MUltiple Stimuli with Hidden Reference and Anchor (множинні стимули зі схованим еталоном і якірними точками);

AMPS – Advanced Mobile Phone System (просунута система мобільного зв'язку);

GSM – Global System for Mobile Communications (глобальна система мобільного зв'язку);

GMSK – Gaussian Minimum Shift Keying (гауссівська мінімальна зсувна маніпуляція);

TDMA – Time Division Multiple Access (множинний доступ з часовим поділом);

FDMA – Frequency Division Multiple Access (множинний доступ з частотним поділом);

AES – Advanced Encryption Standard (просунутий стандарт шифрування);

RSA – Rivest-Shamir-Adleman (шифр RSA);

p2p – Peer-to-Peer (однорангові мережі);

GPRS – General Packet Radio Service (загальна пакетна радіослужба);

EDGE – Enhanced Data rates for GSM Evolution (підвищені швидкості передачі даних для еволюції GSM);

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

FR – Full Rate (повна швидкість);

EFR – Enhanced Full Rate (підвищена повна швидкість);

AMR – Adaptive Multi-Rate (адаптивний багаторазовий швидкісний режим);

MFCC – Мел-частотний кепстральний коефіцієнт;

IDCT – Inverse Discrete Cosine Transform (обернене дискретне косинусне перетворення);

ВСТУП

Передмова

Захист інформації був і залишається важливим питанням забезпечення інформаційної та фізичної безпеки.

В останні роки це важливе питання також знов набуло актуального характеру.

Одним із напрямків досліджень, які продовжують проводитися в цій області є захист голосових комунікацій через мобільні системи зв'язку і в першу чергу — через мережу GSM, як найбільш поширену і таку, яка покриває значні території та використовується в життєдіяльності людей.

Незважаючи на те, що система GSM нібито передбачає шифрування даних на канальному рівні, головним постулатом теорії захисту інформації є вживання відповідних заходів захисту інформації за схемою p2p без довіри до будь-яких інших сторонніх сервісів.

Ефективне шифрування інформації забезпечується лише перевіреними криптографіяними методами. Для цього первісний мовленнєвий сигнал має перетворюватися в цифрову форму, після чого цифрові дані мають шифруватися. Оскількі канал, через яких треба передавати інформацію є аналоговий, то після шифрування цифрових даних вони мають перетворювати на аналогивий сигнал, здатний проходити через відповідний аналоговий канал, тобто має застосовуватися певний спосіб цифрової модуляції.

В системі GSM застосовується параметричне перетворення мовленнєвих сигналів на цифрові параметри, потік яких передається між абонентськими терміналами та базовими станціями. Параметричні кодеки оптимізовані саме для роботи з мовленнєвими сигналами і сильно спотворюють сигнали іншої природи. Цим пояснюється те, що реалізація вказаного підходу до захисту мовлення для систем типу GSM є ускладненою, і в першу чергу тим, що такі канали зв'язку є нестаціонарними, нелінійними і параметричними, крім того не призначені для

роботи з сигналами, які не схожі на мовленнєві за структурую та характеристиками. Тому пропустити через канал сигнал зі стандартною цифровою модуляцією є дуже складно.

При великих навантаженнях на базові станції параметричні кодеки абонентських терміналів переводяться базовою станцією в режими використання менших мережевих ресурсів, — на менші швидкості передачі потоків параметрів мовлення, що приводить до більш сильних спотворень сигналів на вході і виході каналу. Переключення на низкошвидкісні режими роботи також спостерігається через адміністративне втручання в роботу системи — в особливих умовах адміністратор каналів навмисно переводить їх в низько швидкісний режим для ускладнення реалізації будь-яких систем шифрування за схемою p2p, а всі цифрові сервіси відключалися. Такі дії ґрунтуються на тому, що сучасні методи кодування мовлення не можуть забезпечити швидкості потоків менше 600…300 bps, а реалізувати модульовану передачу навіть таких потоків через параметричні нелінійні нестаціонарні канали GSM при низькорежимному кодуванні на вході і виході каналу майже неможливо.

Одним з рішень цієї проблеми є дослідження можливості застосування систем Speech-to-Text та Text-to-Speech замість вокодерів на вході p2p шифратора та дешифратора системи захищеного голосового зв’язку.

Саме така задача і висувається в даній роботі.

Сутність наукової задачі

Попередні пошукові дослідження показали, що стан справ в області автоматичного розпізнавання мовлення, переведення його в тест або псевдо текст та синтез квазімовленевих сигналів за таким описом дає можливість розглядати їх як конкурентів класичних вокодерних систем, тобто досліджувати можливості їх застосування в режимі реального часу при потоковій наскрізній обробці сигналів.

При цьому ставиться задача як дослідження самих систем STT та TSS, їх характеристик роботи при побудові речень зі слів декількох мов одночасно, так і

можливості їх роботи в квазі кадровому поточному режимі, оцінювання погіршення точності розпізнавання при цьому, величин затримки та якості синтезу.

Актуальність теми

В даній роботі розглядається один зі шляхів побудови захищеного голосового зв'язку через голосовий канал GSM, що відповідає сучасним потребам оборонного характеру.

Зв'язок роботи з науковими програмами, планами, темами

Дана робота відповідає тематиці інклюзивних робіт, які проводяться на кафедрі радіоінженерії НТУУ "КПІ" по дослідженню і розробці захищених систем голосового зв'язку та тематиці інноваційних робіт в рамках завдань оборонного характеру із зовнішнім інвестором.

Мета і задачі досліджень

Метою досліджень є отримання інформації про можливість застосування систем STT та TSS для наскрізного потокового кодування багатомовних мовленнєвих сигналів в реальному часі, можливість інтеграції таких систем в систему передачі шифрованого мовлення через голосові низькошвидкісні канали GSM, можливість реалізації таких систем на базі портативних мікроелектронних пристроїв, оцінювання якісних і кількісних характеристик роботи таких систем в означених умовах.

Головні задачі дослідження

Головною задачею дослідження є отримання результатів щодо можливостей та шляхів оптимізації алгоритмів розпізнавання та синтезу мовлення з метою зменшення потрібної швидкості для передачі квазітекстового їх опису при збереженні розбірливості синтезованих сигналів, для чого необхідно:

– дослідити особливості роботи систем STT та TSS,

– дослідити можливості застосування систем STT та TSS для багатомовних мовленнєвих сигналів,
– провести оптимізацію налаштувань складових частин STT та TSS для забезпечення їх роботи з різними мовами одночасно та для зменшення затримки розпізнавання,
– провести експериментальні випробовування наскрізної системи STT та TSS при її реалізації на портативних мікроелектронних пристроях.

Об’єкт досліджень — процеси перетворення сигналів в голосовому каналі системи GSM; процеси параметричного перетворення мовленнєвих сигналів в вокодерних системах та системах speech-to-text та text-to-speech; методи модуляції даних для їх передачі через голосові канали системи GSM.

Предмет досліджень — Системи speech-to-text та text-to-speech, можливість їх наскрізного потокового застосування в реального часі з мінімальними наскрізними затримками для забезпечення роботи з різними мовами, можливості криптографічного p2p шифрування псевдотексту, можливості перетворення псевдотексту на speech-like сигнали, можливості передачі speech-like сигналів через голосовий канал системи GSM при використанні кодеків AMR-4750 з мінімальною швидкістю передачі параметричних даних.

Методи досліджень — методи комп’ютерного моделювання процесів цифрової обробки сигналів, методи проведення натурних лабораторних експериментів.

Наукова новизна отриманих результатів

Новими теоретичними результатами, отриманими автором магістерської дисертації разом з керівником є:

– методика побудови наскрізної системи STT та TSS для потокового мовлення,

- результати випробовування наскрізної системи STT та TSS при потоковому розпізнаванні багатомовних сигналів,
- результати реалізації наскрізної системи STT та TSS на портативних мікроелектронних пристроях у вигляді експериментальної моделі,
- результати випробовування експериментальної моделі наскрізної системи STT та TSS в реальному часі.

Практичне значення отриманих результатів

Практичними (прикладними) результатами магістерської дисертації, отриманими автором самостійно є:

- адаптоване ПЗ та методика побудови наскрізної системи STT та TSS на портативних мікроелектронних пристроях,
- сама експериментальна модель наскрізної системи STT та TSS.

Особистий вклад магістранта

Всі практичні результати досліджень, пов’язані з розробкою і реалізацією методики побудови наскрізної системи STT та TTS, а також дослідженням її роботи виконувалися автором самостійно.

Апробація результатів роботи

Результати досліджень, що наведені в дипломній роботі будуть оприлюдненими під час доповідей на науковому семінарі фізико-математичної школи МННЦ ІТіС МОНУ в с. Жукін в червні-липні 2024 р.

Публікації та доповіді на конференціях

Результати, які отримані під час виконання дисертаційної роботи увійдуть в навчальний посібник до відповідної дисципліни для магістрів РТФ.

1. АНАЛІЗ ПРОБЛЕМ ТА ПОШУК ОПТИМАЛЬНОЇ КОНФІГУРАЦІЇ

1.1 Передумови проблеми

Існує багато виробників, які розробляють та продають продукти для безпечних систем мобільного зв'язку та високошвидкісного шифрування для телекомунікаційних ліній і ліній передачі даних. В Україні такі розробки почали з’являтися на початку 1990-х років, наприклад компанія Орех виготовляла пристрої, в яких мовлення шифрувалося методом частотно-часових перестановок та передавалося аналоговими каналами зв’язку, а пізніше —початковий мовленнєвий сигнал перетворювався на цифровий, здійснювалося його параметричне кодування за алгоритмом CELP, рис. 1.1, кодовані параметри шифрувалися та передавалися через канали ТЧ з використання модему V32. Українська компанія SayCom в середині 1990-х років виготовляла пристрої, в яких мовлення також піддавалося параметричному перетворенню за алгоритмом CELP, вихідний потік цифрових даних шифрувався та передавався через мережу GSM (глобальної систему мобільного зв’язку) з використанням сервісу CSD (Circuit Switched Data або службу даних з комутацією каналів — так званий «дзвінок данних»).

Рисунок 1.1 – Спрощена структурна схема кодера (a) та декодера (b) за алгоритмом CELP.

В якості іншого прикладу можна згадати компанію Sectra Communications AB, яка з середини 1990-х років розробила серію продуктів під назвою Tiger для особистого спілкування, — портативних пристроїв, які забезпечували шифрування мовлення та передачу цифрових даних службою CSD GSM, рис. 1.2.

Рисунок 1.2 – Приклад системи передачі шифрованого мовлення з використанням служби CSD GSM

Самі пристрої Tiger не містили модуля GSM; натомість він підключався до стільникового телефону GSM через Bluetooth, щоб отримати послугу CSD. На рис. 1.2 показано, як два пристрої з’єднуються за допомогою стільникових телефонів.

1.2 Аналіз сучасних цифрових систем голосового зв'язку

1.2.1 Порівняння аналогових і цифрових систем

Розглянемо основні технічні аспекти та переваги кожної з цих технологій.

Аналогові системи використовують безперервні сигнали для передачі голосу. Амплітуда або частота аналогового сигналу змінюється відповідно до амплітуди або частоти голосового сигналу. Наприклад, в системах AMPS використовувалась аналогова передача даних. Цифрові системи, навпаки,

перетворюють голосовий сигнал в цифровий, використовуючи процеси кодування та модуляції. В найпростіших системах голосовий сигнал спочатку оцифровується з використанням PCM, а потім перетворюється на цифровий потік, який модулюються з використанням певного виду цифрової модуляції, після чого модульований сигнал передається через канал зв'язку. В більш ефективних системах використовується параметричне кодування мовлення, рис. 1.3, яке дозволяє суттєво зменшити швидкість потоку цифрових даних і зменшити вимоги до каналу зв'язку. Прикладом є система GSM, яка використовує кодування мовлення на базі лінійного прогнозування.

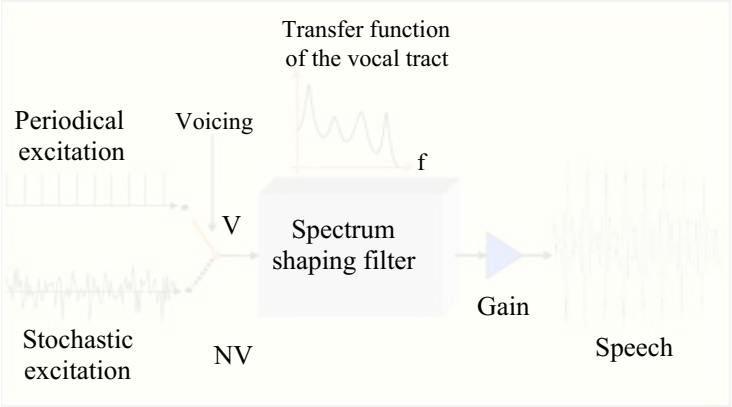

Рисунок 1.1 – Принцип синтезу мовленнєвого сигналу при параметричному його кодуванні за *DSP C*5000, *TI*

Якість звуку в аналогових системах залежить від рівня перешкод і шумів на лінії. Відстань до базової станції також впливає на якість зв'язку, оскільки з ростом відстані сигнал слабшає. Завдяки цифровій обробці сигналу, якість звуку в цифрових системах залишається стабільною і високою незалежно від відстані до базової станції. Цифрові методи, такі як кодек RPE-LTP в GSM, забезпечують високу чіткість і зрозумілість мови навіть в умовах слабкого сигналу.

Аналогові системи використовують один канал на одного користувача, що призводить до низької ефективності використання спектру. Це обмежує кількість одночасних користувачів в одній зоні покриття. Цифрові системи використовують

методи множинного доступу, такі як TDMA та FDMA, що дозволяє одночасно обслуговувати більшу кількість користувачів на одному частотному діапазоні. Наприклад, в GSM кожен частотний канал поділяється на часові слоти, що дозволяє використовувати один канал багатьма користувачами.

Такі системи мають низький рівень безпеки, оскільки аналогові сигнали легко перехопити та розшифрувати. Наприклад, в AMPS перехоплення сигналу було відносно простим завданням. Цифрові системи використовують методи шифрування для захисту даних. У GSM, наприклад, використовується алгоритм шифрування A5/1, що забезпечує високий рівень безпеки і ускладнює перехоплення і розшифровку сигналу.

Аналогові системи обмежуються переважно голосовими послугами. Передача даних або додаткові сервіси, такі як SMS або інтернет, є складними для реалізації і мають низьку ефективність. Цифрові системи підтримують широкий спектр додаткових послуг, включаючи SMS, MMS, передачу даних, інтернет-доступ, відеозв'язок та багато інших. Цифрові системи легко інтегруються з іншими технологіями і сервісами, забезпечуючи універсальність і гнучкість використання.

Аналогові системи мають відносно просту інфраструктуру, але більш вразливі до збоїв і перешкод. Надійність зв'язку може знижуватися через вплив фізичних факторів, таких як погода або природні перешкоди. Цифрові системи мають складнішу інфраструктуру, включаючи цифрові комутатори, базові станції та мережеві елементи. Проте, завдяки сучасним технологіям управління мережею, цифрові системи забезпечують високу надійність зв'язку і стійкість до зовнішніх впливів.

1.2.2 GSM: Технічний Аналіз

GSM використовує GMSK як основну схему модуляції. GMSK є формою частотної модуляції, яка забезпечує високу ефективність спектру та мінімальні перешкоди між каналами. Функціонує на різних частотних діапазонах, таких як 900 MHz, 1800 MHz та 1900 MHz.

GSM використовує комбінацію технологій TDMA та FDMA для ефективного розподілу каналів зв'язку між користувачами.

Архітектура GSM складається з декількох основних підсистем:

- MS (Mobile Station): Мобільний телефон або інший пристрій користувача, який забезпечує голосовий зв'язок та передачу даних.
- BTS (Base Transceiver Station): Базова станція, яка здійснює радіозв'язок з мобільними станціями в зоні покриття.
- BSC (Base Station Controller): Контролер базових станцій, який управляє кількома BTS, забезпечуючи комутацію викликів та управління ресурсами.
- MSC (Mobile Switching Center): Мобільний комутаційний центр, що забезпечує з'єднання між абонентами та з іншими мережами.
- HLR (Home Location Register): База даних, що містить інформацію про абонентів, включаючи їх місцезнаходження та послуги.
- VLR (Visitor Location Register): Тимчасова база даних, яка зберігає інформацію про абонентів, що перебувають у зоні покриття MSC.

Рисунок 1.2 – робота мережі GSM [9]

Однією з основних переваг GSM є його широке покриття та сумісність між операторами, що дозволяє користувачам мати стабільний зв'язок у більшості регіонів світу. Висока якість зв'язку та низький рівень перешкод досягаються завдяки цифровій обробці сигналу та ефективному використанню спектру. Крім того, GSM підтримує широкий спектр додаткових послуг, включаючи SMS, MMS, передачу даних та інтернет-доступ, що робить цю систему універсальною і гнучкою у використанні.

1.3 Проблема захисту інформації

Захист інформації у цифрових системах зв'язку стикається з кількома основними викликами:

- Перехоплення даних. Зловмисники можуть перехоплювати інформацію, що передається через мережу, використовуючи різноманітні методи, такі як сніффінг або прослуховування.
- Незаконний доступ. Може відбуватися несанкціонований доступ до конфіденційних даних через вразливості в системах безпеки.
- Маніпуляція даними. Дані можуть бути змінені або підроблені під час передачі, що може призвести до викривлення інформації та втрати її цілісності.
- Атаки з відмовою в обслуговуванні (DoS). Зловмисники можуть здійснювати атаки, спрямовані на зниження доступності сервісів через перевантаження мережі.

Для вирішення проблем захисту інформації використовуються різноманітні методи, кожен з яких має свої особливості та ефективність.

Шифрування є основним методом захисту даних. Використання криптографічних алгоритмів дозволяє перетворювати інформацію у такий спосіб, що вона стає недоступною для зловмисників без відповідного ключа. Існують різні типи шифрування: симетричне і асиметричне. Симетричне шифрування використовує один ключ для шифрування та дешифрування даних. Прикладом є алгоритм AES. Асиметричне шифрування використовує пару ключів - публічний і

приватний. Публічний ключ використовується для шифрування, а приватний - для дешифрування. Прикладом є RSA.

Використання файрволів також є ефективним способом захисту інформації. Файрволи допомагають контролювати доступ до мережі, блокуючи небажані з'єднання.

Захист цілісності даних є ще одним важливим аспектом інформаційної безпеки. Для забезпечення цілісності даних використовуються хеш-функції, які дозволяють виявити будь-які зміни в інформації. Прикладом є алгоритми MD5, SHA-1, SHA-256 тощо. Ці алгоритми генерують унікальний хеш-код для кожного набору даних, що дозволяє легко виявити будь-які зміни або підробки даних.

P2P криптографічний захист стає все більш актуальним у сучасних цифрових системах зв'язку, забезпечуючи максимальну безпеку і конфіденційність переданих даних. Основна перевага P2P криптографії полягає у відсутності централізованого сервера, що мінімізує ризик компрометації всієї системи через злом одного елемента.

Основні переваги P2P криптографічного захисту :

- Високий рівень безпеки. Використання складних криптографічних алгоритмів забезпечує надійний захист даних від перехоплення та дешифрування.
- Конфіденційність. Всі передані дані шифруються на стороні відправника і розшифровуються тільки на стороні одержувача, що забезпечує максимальну конфіденційність.
- Надійність та стійкість до атак. Відсутність центрального сервера знижує ризик DoS атак і підвищує стійкість системи до збоїв.

1.4 Проблеми використання GSM

Однією з основних проблем GSM є негарантованість працездатності каналів цифрової передачі даних в особливих умовах. Високоякісні режими передачі

даних, такі як GPRS та EDGE, можуть стикатися з труднощами в складних умовах експлуатації. Ці умови включають:

- Переповненість мережі. У пікові періоди, коли велика кількість користувачів одночасно підключаються до мережі, швидкість передачі даних може суттєво знизитися.
- Віддалені або сільські райони. У місцевостях з низькою щільністю базових станцій сигнал може бути слабким, що призводить до зниження швидкості передачі даних та втрат пакетів.
- Залізобетонні будівлі або підземні приміщення. Сигнал GSM важко проникає через товсті стіни або під землю, що також впливає на якість зв'язку та передачу даних.

Ці фактори можуть призвести до нестабільної роботи каналів передачі даних, що особливо критично для додатків, які вимагають високої швидкості та надійності передачі даних. Голосовий канал в GSM також має свої специфічні проблеми, пов'язані з параметричністю, нелінійністю, нестаціонарністю та негарантованістю працездатності високоякісних режимів кодування мовлення.

Голосовий канал в GSM має параметричний характер. Це обумовлено тим, що в системі GSM (а також в інших системах голосового зв’язку) мовленнєвий сигнал розбивається на кадри, після чого кожен кадр сигналу параметрично перетворюється голосовими кодерами в певний набір параметрів керування відповідною моделлю синтезатора (при цьому кодується форму спектральної обвідної, сигнал збудження, період основного тону, та підсилення). Далі квантовані версії параметрів кожного кадру передаються через канал зв’язку у вигляді індексів відповідних кодових книг. На приймальній стороні за прийнятими індексами з кодових книг відновлюються квантовані значення параметрів, які підставляються в модель синтезатора.

В системі GSM застосовуються різні кодери з різною якістю кодування мовлення і швидкість передачі параметрів через мережу, такі як FR, EFR, HR, AMR, табл. 1.1.

Таблиця 1.1 - Характеристики параметричного кодування мовлення в кодера різних систем голосового зв'язку
(за матеріалами дисертаційної роботи Павлова О.І.)

Кодер МС	Тривалість кадру, мс	Параметри аналізу/синтезу МС та структура фільтру, яким визначається ФСО	СОЧисло параметрів для	Параметри, які квантуються та тип квантування	олюютьсяПараметри, які	оляціїКількість сегментів	Параметри, індекси або я значення яких	етрівЧисло біт на вектор	Число біт на 1 параметр за 10 мс
GSM FR (RPE-LTP)[1]	20	Драбинна (решітчата), КВ	8	ЛВП, скалярне	ЛВП	4	ЛВП	36	2,25
GSM EFR (ACELP)[2]	20	Пряма, КЛП, 2 рази на кадр	10	ЛСЧ, скалярне	ЛСП	2	ЛСЧ	38	1,90
GSM HR (VSELP)[3]	20	Драбинна (решітчата), КВ, 1 раз на кадр	10	КВ, векторне, 3-сегменти	КЛП	4	КВ	28	1,40
GSM AMR (ACELP)[4]	20	Пряма, КЛП, 2 рази на кадр	8	ЛСЧ, векторне, розщеплене	ЛСП	4	ЛВП	27	23–1,44—1,69
TIA IS-54 (VSELP)[5]	20	Драбинна (решітчата), КВ	10	КВ, скалярне	КЛП	4	КВ	38	1,90
CDMA EVRC (RCELP)[6]	20	Пряма, КЛП	10	ЛСЧ, векторне, розщеплене	ЛСЧ	4	ЛСЧ	8—28	0,4—1,40
TETRA (ACELP)[7]	30	Пряма, КЛП	10	ЛСП, векторне, розщеплене	ЛСП	4	ЛСП	26	0,86
FS-1015, *(LPC-10)*[8]	22,5	Драбинна (решітчата), КВ	10	ЛВП та КВ, скалярне	—	—	ЛВП та КВ	41	1,82
STANAG 4591 *(FS MELP)*[9]	22,5	Пряма, КЛП	10	ЛСЧ, векторне, 4-каскадне	ЛСЧ	4	ЛСЧ	25	1,11
FS-1016, *(CELP)*[10]	30	Пряма, КЛП	10	ЛСЧ, скалярне	ЛСЧ	4	ЛСЧ	34	1,13
ITU-T G.729, (CS-ACELP)[11]	10	Пряма, КЛП	10	ЛСЧ, векторне, комбіноване	ЛСП	2	ЛСЧ	18	1,80

МС — мовленнєвий сигнал, ФСО — форма спектральної обвідної, КВ — коефіцієнти відбиття, КЛП — коефіцієнти лінійного прогнозування, ЛВП — логарифми відношення площ, ЛСЧ — лінійні спектральні частоти, ЛСП — лінійні спектральні параметри,

Параметрична робота таких кодерів оптимізована під особливості мовленнєвих сигналів і не забезпечує неспотворену передачу через канал GSM сигналів, які традиційно використовуються для цифрової модуляції при передачі потоків цифрових даних (в нашому випадку — цифрових потоків p2p

1 ETSI GSM-FR 06.10 RPE-LTP v5.1.1, 1998 (ETSI 300 961)

2 ETSI GSM-EFR 06.60 ACELP v5.0.0, 1996 (ETSI 300 726)

3 ETSI GSM-HR 06.20 VSELP v4.2.1, 1995 (ETSI 300 581-2)

4 ETSI GSM-AMR 06.90 ACELP v7.2.1, 1998 (ETSI 301 704)

5 Theory and Implementation of the Digital Cellular Standard Voice Coder VSELP on the TMS320C5x (spra136)

6 3GPP2 C.S0014-A Enhanced Variable Rate Codec v1.0, 2004

7 ETSI TETRA FR Speech codec ACELP Part 2 v1.3.1, 2005 (ETSI EN 300 395-2)

8 U.S. Department of Defense. LPC-10 2400 bps Voice Coder. Release 1.0. October 1993

9 STANAG 4591 C3 (Ed.1) - The 600 bit/s, 1200 bit/s and 2400 bit/s NATO Interoperable Narrow Band Voice Coder

10 FED-STD-1016-CELP, 1991. Wai C. Chu. Speech Coding Algorithms. Foundation and Evolution of Standardized Coders. // Published by John Wiley & Sons, Inc., Hoboken, New Jersey. Published simultaneously in Canada. — 2003.

11 ITU-T Recommendation G.729.

шифрованого мовлення). Гармонічні коливання, прості стандартні сигнали типу послідовності прямокутних чи трикутних імпульсів, коливання з АМ, ЧМ, ФМ та інші не можуть бути переданими без суттєвих спотворень через такі параметричні канали.

Існують спроби розв'язати цю проблему за допомогою застосування Speech-Like модемів, — пристроїв, які перетворюють потік цифрових даних на Speech-Like сигнал, який би з більшою мірою успішності пройшов би через параметричний голосовий канал GSM, а квазімовленевий сигнал на виході синтезатора приймача міг би бути перетвореним на первісний потік цифрових даних з найменьшими похибками.

Рисунок 1.3 - Speech-Like модуляція при передачі даних через голосовий канал комунікаційної мережі

В цьому випадку корисна швидкість передачі цифрових даних від шифратора на вході Speech-Like модему є значно меншою за швидкість передачі параметрів кодування мовлення між терміналом і базовою станцією GSM

В реальній системі GSM крім параметричного кодера в передавальному терміналі та параметричного декодера в приймальному терміналі можуть бути

додаткові параметричні перетворення сигналів, пов’язані з тим, що передавальний і приймальний термінали підключені не до однієї тої самої базової станції, а до різних. В цьому випадку треба враховувати передачу сигналів між базовими станціями, яка може здійснюватися як в цифровому вигляді (наприклад, через магістральні радіоканали), так і через канали PSTN. В останньому випадку буде мати місце додаткове параметричне перетворення GSM-to-PSTN на вході в канал PSTN, та додаткове параметричне перетворення PSTN-to-GSM на виході з каналу PSTN та вході в канал GSM, рис. 1.6. Ясно, що може бути декілька таких перетворень і перетворень, пов’язаних з тим, що термінальні пристрої можуть працювати в різних режимах кодування через різну навантаженність на їх базові станції, рис. 1.7.

Рисунок 1.4 - Приклад додаткових перетворень сигналів в системі GSM при включенні в канал зв’язку мережи PSTN

Рисунок 1.5 - Приклад додаткових перетворень сигналів в системі GSM при різних режимах роботи голосових кодексів в початковому і кінцевому терміналах

Крім особливостей параметричного перетворення сигналів треба враховувати також і залежність якості передачі сигналів від таких факорів, як наявність шумів, інтерференції та затримки сигналів. Наприклад, стандартне співвідношення сигнал/шум (SNR) для забезпечення прийнятної якості голосу в GSM становить приблизно 20 дБ. Це співвідношення може змінюватися в залежності від умов середовища, що впливає на стабільність та якість зв'язку.

Нелінійність голосового каналу в GSM може призводити до спотворення голосових сигналів. Це викликано тим, що різні частотні компоненти сигналу можуть передаватися з різними амплітудами та фазами, що викликає нелінійні спотворення. Для мінімізації таких спотворень GSM використовує методи корекції помилок, такі як код Ріда-Соломона та згортальне кодування.

Нестаціонарність голосового каналу означає, що характеристики каналу можуть змінюватися з часом. Наприклад, рух користувача зі швидкістю 30 км/год може призводити до доплерівських зсувів частоти, що впливає на якість передачі сигналу. Нестаціонарність каналу ускладнює забезпечення стабільної якості зв'язку, оскільки система повинна постійно адаптуватися до змін умов.

Голосовий канал в GSM підтримує кілька режимів кодування мовлення, таких як FR з бітрейтом 13 кбіт/с, EFR з бітрейтом 12,2 кбіт/с та AMR, який може змінювати бітрейт від 4,75 до 12,2 кбіт/с в залежності від умов зв'язку, табл. 1.2.

Таблиця 1.2

Table 1: Bit allocation of the AMR coding algorithm for 20 ms frame

Mode	Parameter	1st subframe	2nd subframe	3rd subframe	4th subframe	total per frame
12.2 kbit/s (GSM EFR)	2 LSP sets					38
	Pitch delay	9	6	9	6	30
	Pitch gain	4	4	4	4	16
	Algebraic code	35	35	35	35	140
	Codebook gain	5	5	5	5	20
	Total					244
10.2 kbit/s	LSP set					26
	Pitch delay	8	5	8	5	26
	Algebraic code	31	31	31	31	124
	Gains	7	7	7	7	28
	Total					204
7.95 kbit/s	LSP sets					27
	Pitch delay	8	6	8	6	28
	Pitch gain	4	4	4	4	16
	Algebraic code	17	17	17	17	68
	Codebook gain	5	5	5	5	20
	Total					159
7.40 kbit/s (DAMPS EFR)	LSP set					26
	Pitch delay	8	5	8	5	26
	Algebraic code	17	17	17	17	68
	Gains	7	7	7	7	28
	Total					148
6.70 kbit/s	LSP set					26
	Pitch delay	8	4	8	4	24
	Algebraic code	14	14	14	14	56
	Gains	7	7	7	7	28
	Total					134
5.90 kbit/s	LSP set					26
	Pitch delay	8	4	8	4	24
	Algebraic code	11	11	11	11	44
	Gains	6	6	6	6	24
	Total					118
5.15 kbit/s	LSP set					23
	Pitch delay	8	4	4	4	20
	Algebraic code	9	9	9	9	36
	Gains	6	6	6	6	24
	Total					103
4.75 kbit/s	LSP set					23
	Pitch delay	8	4	4	4	20
	Algebraic code	9	9	9	9	36
	Gains	6		6		16
	Total					95

Однак, працездатність цих режимів не завжди гарантується, рис. 1.8, особливо в умовах високого навантаження на мережу або слабкого сигналу. Високоякісні режими кодування можуть автоматично переключатися на менш якісні, щоб забезпечити збереження зв'язку, але це призводить до зниження якості передачі голосових сигналів.

Codec (used twice)	BER	Standard deviation
GSM FR	28.6%	0.8%
GSM HR	44.4%	0.4%
GSM EFR/AMR 12.2 kbps	1.2%	0.4%
AMR 10.2 kbps	14.3%	0.3%
AMR 4.75 kbps	42.9%	0.5%
AMR-WB 23.85 kbps	0.3%	0.1%
AMR-WB 15.85 kbps	1.6%	0.2%
AMR-WB 12.65 kbps	5.0%	0.8%
AMR-WB 8.85 kbps	23.9%	2.0%
TETRA Codec	34.0%	1.0%

Рисунок 1.6 - Приклад значень BER, який спостерігається при спробах передчі даних з використанням пари однакових голосових кодеків GSM, якы працюють в рызних режимах (за даними дослыджень університету Суррея)

1.5 Створення p2p цифрового каналу через GSM

Створення p2p цифрового каналу всередині голосового каналу GSM є необхідним через кілька причин. По-перше, це дозволяє забезпечити надійну передачу даних в умовах, де традиційні канали передачі даних (наприклад, GPRS або EDGE) не працюють належним чином. По-друге, p2p підхід забезпечує додатковий рівень безпеки, оскільки дані передаються безпосередньо між двома кінцевими користувачами, мінімізуючи ризик перехоплення.

SpeechLike модеми дозволяють передавати цифрові дані через голосовий канал, використовуючи сигнали, схожі на звичайні голосові сигнали. Це робить їх менш помітними для засобів виявлення та дозволяє використовувати наявну голосову інфраструктуру для передачі даних.

Один з варіантів конфігурації системи p2p шифрованого зв'язку з використанням голосових каналів мережи GSM показан на рис. 1.9.

Рисунок 1.7 - Приклад системи шифрованого зв'язку з використанням мережи GSM та PSTN

Для реалізації подібної системи потрібно визначитися з методами кодування мовленнєвих сигналів, які б забезпечили мінімальну швидкість вихідного цифрового потоку при збереженні розбірливості мовлення.

Сучасні вокодери стискають голос зі швидкістю передачі даних від 2000 до 14000 біт на секунду («BPS»). Ентропія або інформаційний вміст людської мови за критерієм розбірливості набагато нижчий (наприклад, від 100 до 200 біт на секунду), що свідчить про те, що майбутні вокодери працюватимуть на набагато нижчих швидкостях.

Наразі існує декілька модемів, доступних для використання зі стільниковою телефонією. Ці модеми використовують звичайні модемні хвилі, які не

відтворюються людським голосовим трактом, і вони надсилають дані зі швидкістю, значно вищою за швидкість ентропії. Це свідчить про те, що ці звичайні модеми використовують переваги неефективності найсучасніших вокодерів і навряд чи працюватимуть належним чином, коли стільникові оператори приймуть нові вокодери з нижчою швидкістю. Розгортання цих звичайних модемів було обмежено з цієї причини [10].

На даний час гарантовано передавати цифрові дані через голосовий канал GSM з кодеком AMR-4750 можна лише на дуже низьких швидкостях. Зокрема, досягнуті швидкості передачі становлять від 75 до 150 біт/с (заяви про досягнення більших швидкостей поки що не мають свого продемонстрованого підтвердження). Ці обмеження обумовлені кількома факторами:

- Низька пропускна здатність. Кодек AMR-4750 має дуже обмежену пропускну здатність, оскільки службова швидкість передачі параметрів у нього найменша серед інших кодеків.
- Високий рівень спотворень. В умовах низької шидкості передачі параметрів між терміналом і базовою станцією GSM спостерігається суттєве зниження якості кодування голосового сигналу (а значить і сигналів Speech-Like модемів), що приводить до сильних спотворень і ускладнює коректну передачу цифрових даних.
- Складність алгоритмів модуляції. Для забезпечення надійної передачі даних необхідно використовувати складні алгоритми Speech-Like модуляції та корекції помилок, що збільшує затримки та вимоги до обчислювальних ресурсів.

1.6 Аналіз можливостей кодування мовлення різними методами

Розбірливість є ключовою вимогою при кодуванні мовлення, оскільки основна мета будь-якої системи передачі голосу — це забезпечення зрозумілості переданого повідомлення. Зниження швидкості передачі даних зазвичай призводить до втрати якості та розбірливості, тому необхідно знайти баланс між швидкістю та якістю.

Вокодери, що працюють на швидкості 600 біт/с — 300 біт/с, часто використовуються в умовах, де необхідно знизити обсяг переданих даних. До таких кодерів можна віднести кодери TWELP 600 bps, 480 bps, 300 bps, MELPe 600 bps, 300 bps, NATO STANG 4591 C3 Eed.1. 600 bps, Codec2 450 bps та інші.

Однак, навіть при цій швидкості, забезпечити розбірливість мовлення в повній мірі не вдається. Незважаючи на те, що такі вокодери можуть зберігати основні характеристики голосу, деталі та нюанси мовлення часто втрачаються, що ускладнює розуміння переданого повідомлення, рис. 1.10.

Рисунок 1.8 - Порівняння значень перцептуальної оцінки якості мовлення для різних мов при використанні різних кодексів з мінімальною швидкістю потоку даних

На рис. 1.10 представлено порівняння значень перцептуальної оцінки якості мовлення (Perceptual Evaluation of Speech Quality, PESQ[12]) для декількох методів її параметричного кодування з використанням лінійного прогнозування при

12

мінімально розумних з точки зору якості швидкостях вихідного потоку параметрів. Видно, що при швидкостях 300 bps якість мовлення для деяких мов при BER 1% зменшується до критичних 2 балів.

Вплив на якість мовлення спотворень, які вносяться в мовленнєвий сигнал параметричними кодеками можна порівняти з впливом на якість мовлення адитивного шуму, наявність якого в каналі приводить до зростання BER при передачі даних і погіршенню PESQ. Такі оцінки представлені на рис. 1.11

Рисунок 1.9 - Порівняння значень перцептуальної оцінки якості мовлення в залежності від відношення С/Ш

Огляд відкритих джерел показує, що на даний час немає доступних кодерів, які можуть забезпечити кодування мовлення зі швидкістю нижче 300 біт/с, зберігаючи при цьому прийнятну розбірливість. Це обумовлено тим, що на таких низьких швидкостях неможливо ефективно передавати всі необхідні для розуміння мовлення частотні компоненти.

Єдиним варіантом, який забезпечує низьку швидкість передачі даних і розбірливість мовлення, є використання систем S2T та T2S. Цей підхід передбачає перетворення мовлення в текстовий формат на стороні відправника S2T і зворотне перетворення тексту в мовлення на стороні одержувача T2S.

Система S2T використовує алгоритми розпізнавання мовлення для перетворення голосового сигналу в текст. Отриманий текстовий файл передається

через низькошвидкісний канал даних, що дозволяє знизити вимоги до пропускної здатності каналу. На стороні одержувача система T2S синтезує мовлення з отриманого тексту, забезпечуючи високу розбірливість.

Основною перевагою цього підходу є можливість забезпечити передачу даних зі швидкістю 75-150 біт/с без втрати розбірливості мовлення. Крім того, використання текстового формату дозволяє застосовувати додаткові методи стиснення даних, що ще більше знижує вимоги до пропускної здатності каналу.

Однак, цей підхід має і свої недоліки. Він вимагає використання складних алгоритмів розпізнавання та синтезу мовлення, що потребує додаткових обчислювальних ресурсів на обох кінцях зв'язку. Крім того, якість та особливості синтезованого псевдомовленнєвого сигналу можуть відрізнятися від таких для оригінального мовлення, що може бути неприйнятним в деяких додатках.

2. СИСТЕМИ TEXT-TO-SPEECH

2.1 Огляд

2.1.1 Визначення TTS та його значення

TTS— це галузь обчислювальної лінгвістики та аудіо-технологій, яка займається перетворенням письмового тексту на усну мову. Використовує спеціальні алгоритми для аналізу тексту та його синтезу в звукову форму, імітуючи людське мовлення.

Основні складові TTS систем включають:

- аналіз тексту (текстовий процесор)
- перетворення на лінгвістичні властивості
- генерацію звуку (аудіо процесор)

2.1.2 Історичний огляд розвитку TTS технологій

Розвиток технологій TTS має багатий історичний контекст, який починається від ранніх механічних спроб імітації людського голосу до сучасних комп'ютеризованих систем.

Основні етапи розвитку:

- 1928. ‘Voder’. Електронний пристрій для синтезу мови
- 1958. ‘Klatt Talker‘. Комп'ютерна система, синтез мови з тексту
- 1984: ‘DECtalk‘. Побутовий пристрій для людей
- 1990-ті. Поява формантного синтезу
- 2000-ті: Розвиток конкатенативного синтезу
- 2010-ті: Глибокоге навчання. Розробка Google's WaveNet та DeepMind's Tacotron
- 2024. Навчання ChatGPT до рівня синтезу мови.

2.2 Основні компоненти TTS системи

2.2.1 Попередня обробка тексту

Попередня обробка тексту відіграє ключову роль у перетворенні письмового тексту на мовленнєвий вивід. Цей компонент системи забезпечує аналіз тексту та його підготовку для подальшого синтезу. Попередня обробка розбита на декілька етапів.

Алгоритм перетворення тексту в фонеми:

- Аналіз тексту: Розбиття тексту на слова та речення для аналізу. Важливо визначити межі слова, що можуть варіюватися залежно від мови.
- Лексичний аналіз: Кожне слово перетворюється на фонеми, базові звукові одиниці, які використовуються для вимови слова. Цей процес може включати спеціальні словники для нестандартного вимовлення.

Розмітка інтонацій та акцентів:

- Процесор інтонацій: Визначає мелодію мови, тон та акценти, що є критично важливим для передачі емоцій та наголосів у мовленні. Робота з інтонацією включає визначення підйомів і падінь тональності голосу, що необхідно для людяності голосу.
- Маркування акцентів: Встановлення акцентів на відповідних складах слова для коректної вимови. Визначення правильних акцентів залежить від контексту речення та може бути складним в полісемантичних або багатозначних умовах.

Додаткові етапи:

- Розбір синтаксичних і граматичних структур: Розпізнавання частин мови, таких як іменники, дієслова, прикметники, що допомагає в коректній інтерпретації тексту.

– Обробка нестандартних та складних текстових елементів: Розпізнавання абревіатур, чисел, спеціальних символів, слів іноземних скорочень.

2.2.2 Синтез мовлення

Синтез мовлення за генерацію звуку на основі вхідних даних, отриманих після попереднього аналізу. Виконує кілька ключових завдань, які спрямовані на створення природного і розбірливого мовлення. Для розмежування алгоритмів поділено на блоки.

Синтез мови:

– Ядро синтезу мови: Після отримання фонем та інтонаційних маркерів, ядро синтезу мови генерує аудіофайли, імітуючи мовлення. Залежно від підходу, може бути використано конкатенативний синтез, де використовуються вже записані мовні фрагменти, або через параметричний синтез, де мова генерується з параметричних моделей.
– Обробка голосу: Включає налаштування висоти, швидкості мовлення та інші звукові ефекти, для покращення результатів.

Генерація звуку:

– Синтез звуків: Ядро системи TTS використовує аудіокарту для створення кінцевого звукового сигналу, що донесеться до користувача.
– Покращення якості звуку: Використання алгоритмів зменшення шумів, збільшення чіткості.

На сьогодні одним з поширених методів синтезу є використання нейронних мереж. Це забезпечує вищу природність та реалістичність мовлення. Також ідуть розробки алгоритмів емоційного забарвлення. Це дозволить системам виражати емоції, змінювати тон, швидкість і інтонацію мовлення ‘на льоту’. Варто не забувати

2.3 Методи синтезу мови

2.3.1 Формантний синтез

Формантний синтез є однією з технік синтезу мови, яка базується на моделюванні вокального тракту та його резонансних властивостей, відомих як форманти. Цей метод використовується для імітації звучання людського мовлення шляхом створення звуків, що відповідають основним резонансам голосового тракту.

Основні принципи формантного синтезу:

- Моделювання вокального тракту: Використовуючи математичні формули та алгоритми, формантний синтез відтворює активність вокального тракту людини під час мовлення. Це включає моделювання різних частин вокального тракту.
- Регуляція формантів: Форманти, які є піковими частотами резонансу вокального тракту, моделюються та коригуються для кожного звука. Відповідне налаштування цих формантів дозволяє виробляти різні мовні звуки, що імітуються у відповідності до норм людської мови.

На основі налаштувань формантів та інших параметрів мовлення, таких як інтенсивність та тон, згенерований звук виходить через звукові модулі, створюючи мовлення.

Найпростіший формантний синтезатор – каскадний. Складається із послідовно з'єднаних смугових резонаторів, де вихід кожного резонатора подається на вхід наступного. Для керування каскадною структурою потрібні лише формантні частоти. Основною перевагою каскадної структури є те, що відносні амплітуди формант для голосних не потребують окремого регулювання [2].

Рисунок 2.1 - Базова структура каскадного синтезатора формант [2]

Також існує інший базовий варінат - синтезатор паралельної форманти. Складається з паралельно з'єднаних резонаторів. Сигнал збудження подається на всі форманти одночасно, а їхні виходи підсумовуються. Вихідні сигнали формантних резонаторів повинні підсумовуватися в протилежній фазі, щоб уникнути небажаних нулів або антирезонансів у частотній характеристиці [3]. Паралельна структура дозволяє окремо контролювати пропускну здатність і посилення для кожної форманти, що потребує додаткової контрольної інформації.

Рисунок 2.2 - Базова структура синтезатора паралельних формант [2]

Наверені вище структури мають плюси і мінуси. Тому було розроблено і запантентовано більш складний формантний синтезатор. Він включає як каскадний, так і паралельний синтезатори з додатковими резонаторами та антирезонаторами для назалізованих звуків [4]. Запатентовано схему для реалізації пристрою SYNTE3.

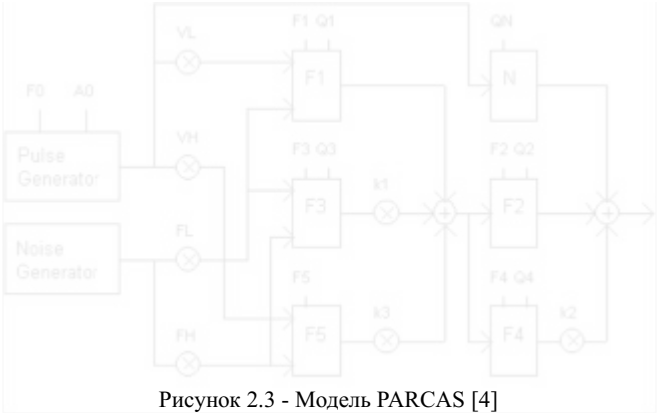

Рисунок 2.3 - Модель PARCAS [4]

Використання формантного синтезу дозволяє детально контролювати характеристики мовлення при низькому використанні резурсів. Цей метод не потребує великої обчислювальної потужності або зберігання великих масивів даних

2.3.2 Конкатенативний синтез.

Конкатенативний синтез є одним із найпоширеніших методів синтезу мови у сучасних TTS системах. Базується на з'єднанні вже записаних звукових сегментів мови. Ці сегменти можуть бути різної довжини — від окремих фонем до цілих слів чи фраз.

Одним з головних компонентів системи є база мовних одиниць. Це систематизована, велика база звукових фрагментів мови. Але база самостійно не зможе провести синтез. Для цього використовують програмне забезпечення. Яке аналізує контекст слова, фонетичні характеристики та інші лінгвістичні аспекти, та вибирає найбільш підходящі звукові сегменти.

Розробкою такого програмного забезпечення зайнялась компанія France Telecom. І представила метод PSOLA. Він дозволяє плавно з'єднувати попередньо записані зразки мовлення та контролювати висоту і тривалість звуку. Хоча це не є метод синтезу мовлення, він використовується в комерційних системах синтезу,).

Рисунок 2.4 - Загальна структура конкатенативної системи синтезу [5]

Рисунок 2.5 – принцип роботи PSOLA [6]

Щоб проілюструвати принцип роботи PSOLA, розглянемо наступний базовий алгоритм на рисунку 2.5:

1. Оцінка контуру основної частоти
2. Знаходження висотних періодів. Наприклад, шляхом визначення найбільшого піку в кожному періоді
3. Виділіть вікна мовного сигналу, що охоплюють два періоди висоти тону. Застосуйте віконну функцію
4. Зсуньте вікна так, щоб вони відповідали бажаній довжині періоду.

2.3.3 Синтез на основі параметричних моделей

Синтез на основі параметричних моделей є одним з методів створення мовлення в системах TTS, який використовує математичні моделі для генерації голосового сигналу. Цей метод відрізняється від конкатенативного синтезу, оскільки він не базується на використанні реально записаних звукових сегментів, а створює мову з основних акустичних параметрів мовлення, які генеруються синтетично.

Параметричний синтез включає визначення ключових акустичних параметрів, таких як частота, спектральні характеристики та тембр. Ці параметри моделюються за допомогою алгоритмів машинного навчання або глибокого навчання на основі великих датасетів.

На основі моделей параметрів система генерує аудіосигнал, який імітує мовлення. Технологія використовує спеціальні фільтри та синтезатори для створення фінального звукового потоку.

Рисунок 2.6 – Схематичне зображення системи SPSS [7]

Переваги параметричного синтезу:

- Не потребує зберігання великих обсягів аудіоданих, оскільки звуки генеруються на льоту з параметрів
- Можливість легко маніпулювати параметрами дозволяє модифікувати характеристики мовлення, такі як вік, стать або емоції
- Легше адаптувати до різних мов і діалектів

Недоліки параметричного синтезу:

- Звучить менш природно порівняно з конкатенативним синтезом
- Створення ефективних параметричних моделей вимагає глибоких знань в акустиці, лінгвістиці та машинному навчанні.

2.3.4 Нейронні мережі в TTS.

Використання нейронних мереж для TTS відкрило нові можливості для поліпшення якості та природності мовлення. Нейронні мережі дозволяють моделювати мовлення з високою точністю за допомогою глибокого навчання та великих даних.

Нейронні мережі здатні вивчати нюанси мовлення, такі як інтонація, акцентуація та паузи, а техніки машинного навчання дозволяють системам адаптуватися до стилів мовлення конкретних користувачів або контекстів. Також можуть передавати емоції через тон, швидкість мовлення та інші звукові параметри, що робить TTS системи більш ефективними у випадках, де важлива емоційна взаємодія, наприклад, у віртуальних помічниках або інтерактивних іграх.

Серед сучасних моделей синтезу мовлення виділяються WaveNet і Tacotron. WaveNet, розроблена DeepMind, використовує згорткові нейронні мережі для генерації сирого аудіосигналу, що дозволяє створювати дуже природне мовлення [8]. Tacotron, популярний підхід від Google, перетворює текст на спектрограми, які потім перетворюються у мовлення за допомогою WaveNet або інших вокодерів.

Переваги використання нейронних мереж в TTS:

- Здатність генерувати більш природнє та плавне мовлення
- Легше адаптувати системи до нових мов та акцентів
- Нейронні мережі можуть обробляти широкий спектр мовних варіацій

Виклики та обмеження:

- Потреба у великих обсягах даних та обчислювальних ресурсах для тренування та функціонування.
- Моделі потребують значних зусиль для оптимізації та налаштування.

2.4 Нейронні мережі та глибоке навчання в TTS

2.4.1 Архітектури нейронних мереж, які використовуються у TTS

У сучасних технологіях текст-в-мову (TTS), використання нейронних мереж та глибокого навчання стало ключовим для досягнення високої якості та природності мовлення. Нижче описано дві важливі архітектури, які мають значний вплив на галузь TTS :

- WaveNet - глибока згорткова нейронна мережа, яка генерує звукові хвилі безпосередньо з аудіо семплів. Вона використовує модель авторегресії, де кожен наступний аудіосемпл передбачається на основі попередніх семплів. Архітектура здатна моделювати різні звуки та голоси з високим ступенем деталізації та природності, роблячи її ідеальною для синтезу мовлення. Знайшла використання в Google Assistant. Розробкою займається компація DeepMind.
- Tacotron - end-to-end система синтезу мови, яка перетворює текст прямо в звукові хвилі, використовуючи рекурентні нейронні мережі (RNN). Вона використовує архітектуру, що включає кодер, який аналізує текстові вхідні дані, і декодер, який генерує спектрограми аудіосигналу. Спектрограми

потім перетворюються у звук за допомогою вокодера. Такий алгоритм роботи дозволяє створювати у голосі динамічні інтонації.

Ці архітектури демонструють можливості глибокого навчання в синтезі мовлення, забезпечуючи значні переваги у порівнянні з традиційними методами. Нейронні мережі дозволяють:

- Легко впроваджувати нові мови та акценти без зміни архітектури.
- Можливість налаштовувати голос під конкретні вимоги або сценарії використання.
- Вводити різні емоційні стани у мовлення, що підвищує його природність та ефективність сприйняття.

2.4.2 Переваги використання нейронних мереж.

Нейронні мережі та глибоке навчання внесли значний вклад у технології текст-в-мову (TTS), пропонуючи ряд переваг порівняно з традиційними методами. Проведемо огляд на основні переваги використання нейронних мереж у TTS:

- Підвищена природність мовлення. Нейронні мережі створюють мовлення, яке звучить дуже природно і плавно, передаючи тонкощі людської мови, такі як інтонації, акценти та ритм, що складно досягти традиційними методами.
- Висока гнучкість і масштабованість. TTS-системи на основі нейронних мереж легко адаптуються до різних мов і діалектів, дозволяючи швидко розширюватися на нові ринки. Вони також обробляють різноманітні текстові стилі та складні лінгвістичні структури.
- Емоційна виразність. Сучасні TTS-системи здатні передавати емоції та інтонації, що є важливим для залучення слухачів. Нейронні мережі точно моделюють ці аспекти.

- Контроль над характеристиками голосу. Нейронні мережі дозволяють точно налаштовувати голос, включаючи вік, стать та емоційний тон. Це робить їх ідеальними для застосувань, де необхідно створити певну атмосферу або персоналізацію.
- Ефективність обробки великих даних. Нейронні мережі ефективно обробляють великі набори даних, що дозволяє їм навчатися та вдосконалюватися з кожним новим обсягом вхідної інформації, постійно покращуючи якість мовлення.
- Швидкість впровадження інновацій. Нейронні мережі розвиваються швидше за традиційні технологічні підходи, що дозволяє TTS-системам швидко інтегрувати новітні дослідницькі розробки та вдосконалення.

2.5 Оптимізація мовного синтезу

2.5.1 Методи покращення якості та природності мови.

Оптимізація мовного синтезу в системах текст-в-мову (TTS) спрямована на підвищення розбірливості, природності та емоційної виразності мовлення. Існує кілька ключових методів, які використовуються для досягнення цих цілей.

Один з яких - використання згорткових нейронних мереж (CNN) та рекурентних нейронних мереж (RNN), включаючи LSTM (Long Short-Term Memory) та GRU (Gated Recurrent Units), дозволяє моделювати складні залежності в мовній інформації. Розробка end-to-end моделей, як-от Tacotron та WaveNet, дозволяє безпосередньо перетворювати текст на мовлення, покращуючи природність та розбірливість.

Впровадження алгоритмів для точного моделювання інтонацій, акцентів та ритму мовлення. Це включає аналіз контексту слова у реченні для визначення його емоційного навантаження та синтаксичної ролі, а також розробку технологій, що дозволяють системі адаптувати інтонації залежно від семантичного і прагматичного змісту тексту.

Впровадження механізмів для варіації емоційних станів в мовленні, що може включати зміну темпу, інтонації, та сили голосу. Це підвищує природність сприйняття мовлення та робить взаємодію з TTS системами більш приємною та ефективною. Розвиток алгоритмів для аналізу емоційного змісту тексту та його відтворення через мовлення.

Включення лінгвістичного та семантичного аналізу тексту для розуміння контексту та його впливу на мовлення. Це допомагає TTS системам точніше відтворювати смислові нюанси, зокрема, у складних діалогах або спеціалізованих текстах.

Використання формантного синтезу для регулювання резонансних частот, що відповідають за основні характеристики звучання голосу. Це може значно покращити якість мовлення, роблячи його більш приємним на слух.

2.5.2 Зменшення часу обробки та вимог до ресурсів.

Оптимізація ресурсів та ефективність обробки є критично важливими аспектами при розробці систем TTS, особливо коли вони впроваджуються в мобільні пристрої та вбудовані системи. Розглянемо декілька методів, які використовуються для мінімізації часу обробки та вимог до обчислювальних ресурсів.

Використання ефективніших алгоритмічних структур, таких як оптимізовані згорткові нейронні мережі, дозволяє зменшити кількість необхідних обчислень за рахунок більш ефективного розподілу ресурсів. Крім того, впровадження технік кешування та передобчислення допомагає зберігати результати часто виконуваних операцій, що ще більше знижує навантаження.

Ще одним важливим підходом є розробка прориджених нейронних мереж. У таких архітектурах активні лише невелика частина нейронів на кожному шарі, що

значно знижує обчислювальне навантаження. Цей метод дозволяє зберегти високу продуктивність системи при значно меншому використанні ресурсів.

Квантування та бінаризація ваг також сприяють ефективності обробки. Використання квантованих та бінаризованих нейронних мереж, де ваги та активації представлені меншою кількістю бітів, знижує вимоги до пам'яті та прискорює обчислення. Це особливо корисно для вбудованих систем з обмеженими ресурсами.

Асинхронне виконання дозволяє системам виконувати декілька завдань паралельно або використовувати асинхронне виконання для розподілу навантаження та зменшення часу відгуку. Це підвищує загальну продуктивність та дозволяє ефективніше використовувати наявні ресурси.

Апаратні оптимізації, такі як використання спеціалізованих обчислювальних пристроїв, таких як FPGA або ASIC, значно прискорюють обробку за рахунок апаратної підтримки специфічних операцій. Ці пристрої можуть бути налаштовані на виконання конкретних завдань з високою ефективністю.

Нарешті, хмарні обчислення та крайові обчислення допомагають збалансувати навантаження між локальними пристроями та хмарними сервісами. Це дозволяє зменшити вимоги до локальних обчислювальних ресурсів та прискорити обробку, забезпечуючи високу продуктивність системи навіть на пристроях з обмеженими можливостями.

2.6 Практичне застосування TTS

2.6.1 Розмовні агенти та віртуальні асистенти.

Розмовні агенти та віртуальні асистенти є ключовими застосуваннями технологій TTS, які революціонізували взаємодію між людиною та машиною. Вони інтегровані в побутові пристрої, смартфони, автомобільні системи та численні платформи для надання автоматизованої підтримки, керування задачами, навчання, та розваг.

Основні аспекти використання TTS в розмовних агентах та віртуальних асистентах:

- Інтерактивність. Віртуальні асистенти використовують TTS для надання голосових відповідей на запити користувачів. Це створює враження "живої" бесіди та поліпшує зручність використання.
- Персоналізація. Сучасні системи можуть адаптувати мовний стиль, темп, інтонацію, навіть акцент, відповідно до переваг користувача або контексту запитів.
- Автоматизація обслуговування. Використання TTS дозволяє автоматизувати відповіді на часті питання, забезпечуючи 24/7 підтримку без необхідності участі людського оператора.
- Доступність. Технології TTS допомагають людям з вадами зору або обмеженнями руху, надаючи їм можливість взаємодіяти з технологіями за допомогою голосових команд.
- Інтеграція з іншими технологіями. Віртуальні асистенти часто інтегруються з іншими системами, такими як інтелектуальне освітлення, системи безпеки та медіацентри, що дозволяє керувати ними голосом.

Технічні виклики включають підтримку різних мов та діалектів, розуміння природної мови та контексту розмови, а також забезпечення безпеки та конфіденційності в обробці особистих даних.

Приклади популярних віртуальних асистентів включають Siri від Apple, Google Assistant від Google, Alexa від Amazon та Cortana від Microsoft.

2.6.2 Автоматизація call-центрів.

Автоматизація call-центрів за допомогою технологій TTS та розпізнавання мовлення значно змінила індустрію обслуговування клієнтів. Використання TTS дозволяє компаніям підвищувати ефективність, знижувати витрати та покращувати загальний досвід клієнтів.

Ключові аспекти автоматизації call-центрів:

- Ефективне управління запитами. Системи на базі TTS автоматизують стандартні запити, такі як перевірка рахунків, зміна даних контактів, надання інформації про продукти та послуги, що дозволяє операторам зосередитися на більш складних і важливих завданнях.
- Підвищення доступності. Автоматизовані системи працюють цілодобово, забезпечуючи клієнтам доступ до потрібної інформації у будь-який час без необхідності чекати на з'єднання з оператором.
- Мультимовність. TTS системи можуть працювати з багатьма мовами та діалектами, що розширює базу потенційних клієнтів та підвищує задоволеність неангломовних користувачів.
- Зменшення помилок. Автоматизація знижує кількість людських помилок, які можуть виникати при ручному введенні даних або обробці запитів, забезпечуючи вищий рівень точності обслуговування.
- Інтеграція з іншими системам. Автоматизовані TTS системи можуть інтегруватися з CRM, базами даних та іншими корпоративними системами, покращуючи управління клієнтською інформацією та оптимізуючи взаємодію.

2.7 Проблеми конфіденційності та безпеки

Використання технологій TTS викликає серйозні питання конфіденційності та безпеки, особливо з огляду на зростаючу інтеграцію цих систем у повсякденне життя та бізнес. Є кілька ключових аспектів, які необхідно враховувати.

По-перше, збір та обробка даних є важливим питанням. TTS системи часто збирають великі обсяги голосових даних для тренування та оптимізації. Це викликає питання про те, як і де ці дані зберігаються, хто має до них доступ та як забезпечується їхня безпека.

Зловмисне використання технологій також є потенційною загрозою. Існує ризик використання TTS для створення обманних аудіозаписів або імітації голосів без дозволу осіб. Це може призвести до шахрайства, маніпуляцій та порушення особистої недоторканності.

Крім того, важливо дотримуватися законодавчого регулювання. Наприклад, законодавства, такого як GDPR у Європейському Союзі або CCPA у Каліфорнії, яке регулює збір та використання особистих даних, включаючи голосову інформацію.

Ще одним важливим аспектом є технічні заходи безпеки. Потрібно розробляти та впроваджувати передові методи шифрування та аутентифікації для захисту даних, переданих між пристроями та серверами. Важливо також забезпечити безпеку пристроїв, на яких виконується TTS.

Також користувачі повинні мати чітке розуміння того, як збираються, використовуються та зберігаються їхні голосові дані. Вони також повинні мати можливість легко управляти своїми даними та відкликати згоду на їх обробку.

2.8 Огляд наявних методів оцінювання якості синтезованої мови.

Оцінювання якості синтезованої мови є ключовим елементом у розробці та вдосконаленні систем TTS. Різні методи оцінювання допомагають розробникам зрозуміти, наскільки ефективні їхні системи у створенні природного та зрозумілого мовлення. Основні методи оцінювання можна поділити на дві категорії: об'єктивні та суб'єктивні.

Об'єктивні методи включають кілька підходів. Першим з них є MOS. Це стандартна суб'єктивна техніка, яка часто використовується для оцінювання якості голосу, але тут з певними об'єктивними адаптаціями. Об'єктивний MOS вимірює якість на основі алгоритмічних прогнозів, що корелюють з людськими оцінками. Іншим популярним методом є PESQ, який часто використовується для оцінювання якості мовленевих кодеків і телефонних мереж. Він вимірює відхилення синтезованого мовлення від оригінального запису. Третій метод, VRR, оцінює точність візуального відповідника фонем у синтезованому мовленні, що особливо важливо для систем, що генерують анімаційне мовлення персонажів.

Суб'єктивні методи також відіграють важливу роль у оцінюванні якості синтезованого мовлення. Одним з основних підходів є слухові тести, де учасники оцінюють якість мовлення за різними критеріями, такими як розбірливість, природність і приємність голосу. Ці оцінки відображають враження реальних користувачів і є дуже важливими для розуміння сприйняття мовлення. ABX тести є ще одним методом, де учасникам пропонують прослухати два мовні зразки (A та B) та визначити, який із них ближчий до контрольного зразка X за певними характеристиками. Метод MUSHRA використовується для оцінки мовлення шляхом порівняння декількох варіантів мовлення з прихованим еталонним зразком та анкерним зразком нижчої якості.

Таким чином, використання як об'єктивних, так і суб'єктивних методів оцінювання дозволяє розробникам всебічно оцінити якість синтезованого мовлення та зробити необхідні поліпшення для забезпечення найкращого користувацького досвіду.

3. СИСТЕМИ SPEECH-TO-TEXT

3.1 Вступ

Розпізнавання мови STT - процес перетворення усного мовлення в текст за допомогою комп'ютерних алгоритмів. Технології STT дозволяють машинам "чути" і розуміти людську мову, що є важливим кроком у розвитку інтерфейсів людина-машина.

Значення STT:

- STT значно покращують доступ до інформації для людей з обмеженими можливостями, наприклад, для глухих або слабочуючих осіб. Системи розпізнавання мови можуть автоматично створювати субтитри для відео, що робить контент доступнішим.
- STT дозволяє автоматизувати процеси, які раніше вимагали ручної роботи. Наприклад, транскрибування зустрічей, інтерв'ю або лекцій стає швидшим і точнішим завдяки використанню STT технологій.
- Віртуальні асистенти, такі як Google Assistant, Amazon Alexa або Apple Siri, використовують STT для розуміння команд користувачів. Це дозволяє створювати більш природні та інтуїтивно зрозумілі способи взаємодії з технологіями.
- STT може значно підвищити продуктивність в різних професійних сферах. Наприклад, лікарі можуть диктувати свої примітки, замість того щоб їх писати вручну, що економить час і зменшує ймовірність помилок.
- Інтерактивні голосові відповіді (IVR) та інші системи на основі STT можуть забезпечити більш зручний та ефективний сервіс для клієнтів, зменшуючи час очікування та підвищуючи задоволеність користувачів.

Отже, технології STT мають велике значення у сучасному світі, забезпечуючи нові можливості для комунікації, автоматизації та доступності

інформації. Вони відкривають нові горизонти в різних галузях, від медицини до освіти, роблячи наш світ більш інтерактивним та доступним.

3.2 Основні компоненти STT системи

3.2.1 Обробка аудіосигналу, виділення ознак

Його основна функція полягає в обробці вхідного аудіосигналу та виділенні ключових ознак, які можна використовувати для подальшого розпізнавання мови.

Попередня обробка аудіосигналу включає нормалізацію та фільтрацію шумів. Нормалізація передбачає регулювання рівня гучності сигналу, щоб уникнути впливу дуже тихих або гучних звуків. Фільтрація шумів спрямована на видалення або зменшення фонових шумів, які можуть заважати точному розпізнаванню мови, і може включати застосування методів фільтрації низьких і високих частот.

Фреймування аудіосигналу передбачає розбиття безперервного сигналу на невеликі фрагменти (фрейми), зазвичай тривалістю 20-40 мілісекунд. Це дозволяє аналізувати сигнал у коротких часових інтервалах і виділяти ознаки для кожного фрейму окремо.

Віконування застосовується до кожного фрейму для згладжування країв і мінімізації артефактів, які можуть виникати при фреймуванні. Використовуються віконні функції, такі як Хеммінгова або Хеннінгова вікна.

Перетворення Фур'є застосовується до кожного фрейму для переходу від часової області до частотної. Це дозволяє аналізувати частотний склад сигналу.

MFCC є одним із найпопулярніших методів виділення ознак у STT системах. Спочатку відбувається перетворення сигналу на мел-шкалу за допомогою банку фільтрів, які моделюють людське вухо та виділяють найважливіші частоти. Потім застосовується логарифмічне масштабування до спектру мел-фільтрів для імітації людського сприйняття гучності. Нарешті, використовується зворотне дискретне косинусне перетворення для отримання коефіцієнтів, які представляють кепстральні ознаки.

Перцептивний лінійний передбачувач (PLP) враховує перцептивні аспекти слуху, зменшуючи вплив частот, які менш важливі для людського сприйняття. Це альтернативний метод виділення ознак.

Спектральні коефіцієнти використовуються для додаткового опису ознак сигналу. До них належать спектральна енергія, спектральна площина та інші показники.

Дельта та дельта-дельта коефіцієнти є похідними першого та другого порядку від MFCC або PLP коефіцієнтів. Вони використовуються для захоплення динаміки змін сигналу.

Рисунок 3.1 – Діаграма проходження першого етапу обробки

Загальний порядок виділення ознак:

- Вхідний аудіосигнал
- Попередня обробка (нормалізація, фільтрація шумів)
- Фреймування і віконування
- Перетворення Фур'є
- Виділення ознак (MFCC, PLP, спектральні коефіцієнти)
- Дельта та дельта-дельта коефіцієнти
- Вихідний набір ознак

3.2.2 Перетворення ознак у текст

Перетворення виділених фронтендом ознак у текст. Це включає кілька важливих етапів, таких як моделювання мови, використання акустичних моделей, та обробка текстових даних. Далі розлгянемо декілька популярних варіантів.

Моделі на основі прихованих марковських моделей (HMM) є одними з найпоширеніших для розпізнавання мови. Вони використовують статистичні методи для моделювання послідовностей ознак і їх відповідності фонемам мови. HMM моделюють ймовірності переходів між станами, кожен з яких відповідає певній фонемі або частині фонеми, та емісійні ймовірності, які оцінюють ймовірності того, що певний набір ознак належить до конкретного стану.

Нейронні мережі, такі як рекурентні нейронні мережі (RNN), довгі короткочасні пам'яті (LSTM) та трансформери, стали стандартом для сучасних систем розпізнавання мови (STT). RNN і LSTM здатні обробляти послідовності даних і враховувати контекст, що значно покращує точність розпізнавання мови. Трансформери застосовуються для моделювання довгострокових залежностей у мовних даних, забезпечуючи високу точність та ефективність.

Лексичні моделі, або моделі словників, містять інформацію про вимову слів і відповідність їх фонемам. Це дозволяє системі розпізнавати не тільки окремі фонеми, а й цілі слова, що значно покращує точність розпізнавання.

N-грамові моделі використовуються для моделювання ймовірностей появи певних послідовностей слів. Вони допомагають передбачати наступне слово на основі попередніх кількох (N) слів. Біграми та триграми є найпоширенішими типами N-грамових моделей, які використовують два або три слова для передбачення.

Нейронні мовні моделі, які використовують нейронні мережі для моделювання послідовностей слів, можуть враховувати довші контексти, що покращує точність розпізнавання. Сучасні моделі на основі трансформерів, такі як

GPT (Generative Pre-trained Transformer), використовуються для контекстуалізації та генерації мовних даних.

Рисунок 3.2 – Діаграма проходження другого етапу обробки

Загальний порядок обробки ознак:

- Вхідний набір ознак від фронтенду
- Акустичні моделі (HMM, нейронні мережі)
- Лексичні моделі (словники вимови)
- Мовні моделі (N-грамові, нейронні моделі)
- Декодування (алгоритм Вітербі)
- Постобробка (автоматична корекція, іменний словник)
- Вихідний текст

3.3 Методи розпізнавання мови

3.3.1 Методи на основі шаблонів

Методи розпізнавання мови на основі шаблонів були одними з перших технологій, використаних для перетворення усного мовлення в текст. Вони базуються на порівнянні вхідного мовного сигналу з попередньо збереженими зразками (шаблонами) і знаходженні найбільш відповідного. Розглянемо більш детально ці методи.

Метод на основі шаблонів має такі властивості:

- Збереження шаблонів. Попередньо зберігаються еталонні зразки мовних одиниць (фонем, складів, слів) для кожного мовця або для загального випадку.

- Порівняння з шаблонами. Вхідний мовний сигнал розбивається на фрейми, для кожного з яких виділяються ознаки. Ці ознаки порівнюються з ознаками збережених шаблонів.
- Оцінка схожості. Використовуються різні метрики для оцінки схожості між вхідними ознаками і шаблонами, наприклад, евклідова відстань або кореляційний коефіцієнт.

Динамічне часове вирівнювання (DTW):

- Принцип DTW. Динамічне часове вирівнювання використовується для вирівнювання тимчасових відмінностей між вхідним сигналом і шаблоном. Це особливо корисно для врахування різних швидкостей мовлення.
- Алгоритм DTW. Обчислюється матриця відстаней між фреймами вхідного сигналу і шаблону, після чого знаходиться шлях з мінімальною сумарною відстанню через цю матрицю. Цей шлях вказує на найбільш відповідне вирівнювання між сигналом і шаблоном.

Методи на основі шаблонів мають кілька важливих переваг. По-перше, вони відносно прості у реалізації, не потребують складних математичних моделей, що робить їх доступними для багатьох застосувань. По-друге, ці методи можуть демонструвати високу точність при роботі з обмеженими наборами даних, де мовні варіації є відносно стабільними. Це дозволяє ефективно використовувати їх у ситуаціях з обмеженою кількістю можливих мовних команд.

Однак, методи на основі шаблонів мають і свої недоліки. Їх обмежена масштабованість стає проблемою при роботі з великими словниками або багатьма користувачами, оскільки необхідно зберігати велику кількість шаблонів. Крім того, ці методи чутливі до варіацій мовлення, таких як акценти, інтонації та швидкість мовлення, що може знижувати їх ефективність у реальних умовах. Також вони не мають гнучкості до навчання та адаптації до нових мовних даних

без попереднього оновлення шаблонів, що обмежує їхню здатність до самовдосконалення.

Методи на основі шаблонів знайшли своє застосування у різних сферах. Вони використовувалися у ранніх системах розпізнавання голосових команд, таких як телефонні автоматичні відповідачі або системи голосового управління технікою. Також ці методи знайшли застосування у спеціалізованих пристроях для людей з обмеженими можливостями, де набір команд був обмеженим і фіксованим, що дозволяло досягати високої точності розпізнавання.

3.3.2 Статистичні моделі.

Статистичні моделі відіграють ключову роль у сучасних системах розпізнавання мови, забезпечуючи високу точність і гнучкість у обробці мовних сигналів. Одним з найбільш поширених підходів є приховані марковські моделі (Hidden Markov Models, HMM), які використовуються для моделювання часових послідовностей даних.

Приховані марковські моделі (HMM) є статистичними моделями, що описують систему як послідовність прихованих станів, кожен з яких відповідає певному спостережуваному виходу. У контексті розпізнавання мови, приховані стани відповідають фонемам або групам фонем, а спостережувані виходи — ознакам звукового сигналу.

HMM включає кілька основних компонентів: набір станів, які відповідають певним частинам мовного сигналу; перехідні ймовірності, що визначають ймовірність переходу від одного стану до іншого; емісійні ймовірності, що визначають ймовірність того, що певний стан генерує спостережуваний вихід; і початкові ймовірності, які вказують на ймовірність початку системи з певного стану.

До основних алгоритмів HMM належать: алгоритм Вітербі, що знаходить найбільш ймовірну послідовність станів для даної послідовності спостережень; алгоритм вперед-назад, який обчислює ймовірності станів у середині

послідовності спостережень; та EM-алгоритм (Expectation-Maximization), що використовується для навчання HMM і оптимізації параметрів моделі на основі навчальних даних.

Рисунок 3.3 – приклад приховані марковські моделі

HMM застосовуються для моделювання мовних сигналів у різних контекстах. Для акустичного моделювання HMM використовуються для моделювання фонем або інших мовних одиниць, де кожна фонема моделюється як послідовність станів. У лексичному моделюванні слова моделюються як послідовності фонем, кожна з яких представлена окремою HMM. Для мовного моделювання використовуються статистичні моделі, такі як N-грамові моделі, для визначення ймовірності послідовності слів.

Переваги HMM включають гнучкість у моделюванні часових залежностей та варіацій мовних сигналів, а також наявність добре вивчених алгоритмів для їх навчання та оцінки. HMM також легко інтегруються з іншими статистичними моделями та методами, такими як N-грамові мовні моделі.

Проте HMM мають і свої недоліки. Одним з основних обмежень є припущення про умовну незалежність спостережень за даного стану, що не завжди відповідає реальним мовним сигналам. Крім того, навчання та оцінка HMM можуть вимагати значних обчислювальних ресурсів, особливо для великих словників або тривалих записів.

HMM знаходять застосування у багатьох комерційних системах розпізнавання мови, таких як диктування тексту або автоматичні телефонні системи. Вони також є основою багатьох наукових досліджень у галузі обробки мовних сигналів і стандартним інструментом для багатьох наукових проектів.

3.3.3 Нейронні мережі в STT.

Нейронні мережі, особливо глибокі нейронні мережі (Deep Neural Networks, DNN), стали ключовою технологією в сучасних системах розпізнавання мови (STT). Вони забезпечують високу точність і здатність адаптуватися до різних мовних умов та акцентів.

Нейронні мережі складаються з шарів штучних нейронів, які обробляють вхідні дані. Основні архітектури, що використовуються в STT, включають багатошарові перцептрони (MLP), рекурентні нейронні мережі (RNN) і згорткові нейронні мережі (CNN).

Нейронні мережі навчаються на великих обсягах даних, використовуючи методи зворотного поширення помилки (backpropagation) для оптимізації вагових коефіцієнтів.

Рекурентні нейронні мережі RNN здатні обробляти послідовності даних, зберігаючи інформацію про попередні етапи. Це робить їх особливо корисними для розпізнавання мовлення, де контекст попередніх слів впливає на розпізнавання поточного слова.

Довгі короткочасні пам'яті (Long Short-Term Memory, LSTM) та гейтовані рекурентні одиниці (Gated Recurrent Units, GRU) є вдосконаленнями RNN, що вирішують проблему зникнення градієнта та покращують здатність запам'ятовування довгострокових залежностей.

Згорткові нейронні мережі (CNN) використовують згорткові шари для виділення просторових ознак з вхідних даних. Вони часто застосовуються для попередньої обробки аудіосигналів, виділяючи важливі акустичні ознаки.

CNN ефективно обробляють великі обсяги даних і можуть виділяти складні ознаки, що робить їх корисними для підготовки даних перед подальшою обробкою RNN або трансформерами.

Трансформери використовують механізм самоуваги (self-attention) для обробки послідовностей даних. Це дозволяє їм враховувати контекст на всіх рівнях послідовності, що покращує точність розпізнавання. Моделі, такі як BERT (Bidirectional Encoder Representations from Transformers) і GPT (Generative Pre-trained Transformer), досягли значного успіху в обробці природної мови і застосовуються для задач STT.

Енд-то-енд моделі дозволяють безпосередньо перетворювати аудіосигнал у текст, минаючи проміжні етапи, такі як виділення ознак або фонетичне моделювання. Енд-то-енд підхід спрощує процес обробки і знижує вимоги до ручного налаштування моделі, одночасно підвищуючи точність розпізнавання.

Приклади нейронних мереж у STT:

- DeepSpeech. Модель, розроблена Baidu, використовує глибокі рекурентні нейронні мережі для розпізнавання мовлення.
- WaveNet. Розроблена Google, ця модель генерує звукові сигнали і може використовуватися для синтезу мовлення, а також для покращення якості розпізнавання мовлення.
- Tacotron. Інша модель від Google, яка використовує нейронні мережі для енд-то-енд синтезу мовлення, але її принципи можуть бути застосовані і в задачах розпізнавання.

Переваги нейронних мереж :

- Висока точність. Нейронні мережі забезпечують високу точність розпізнавання завдяки здатності враховувати контекст і обробляти складні залежності.
- Адаптивність. Вони можуть адаптуватися до різних умов мовлення, акцентів та мов.

Проблеми використання:

- Високі вимоги до ресурсів. Навчання і розгортання глибоких нейронних мереж потребує значних обчислювальних ресурсів.
- Необхідність великих даних. Для досягнення високої точності потрібні великі обсяги навчальних даних, що може бути складним завданням для деяких мов або діалектів.

3.4 Нейронні мережі та глибоке навчання в STT

3.4.1 Архітектури нейронних мереж, які використовуються у STT

Нейронні мережі та глибоке навчання здійснили революцію в системах розпізнавання мови (STT), надаючи високу точність і гнучкість. Розглянемо деякі з основних архітектур нейронних мереж, що широко використовуються в STT, включаючи такі моделі, як DeepSpeech і Transformer.

DeepSpeech – це архітектура глибоких нейронних мереж, створена компанією Baidu. Вона використовує глибокі рекурентні нейронні мережі (RNN) для перетворення аудіосигналів у текст. DeepSpeech включає кілька шарів рекурентних нейронних мереж (RNN), зокрема довгі короткочасні пам'яті (LSTM), які можуть зберігати контекстну інформацію протягом тривалого часу. Для навчання моделі використовується функція втрат Connectionist Temporal Classification (CTC), що дозволяє моделі навчатися з необроблених аудіоданих і текстових транскрипцій без необхідності точного вирівнювання між ними. DeepSpeech забезпечує високу точність розпізнавання мовлення і може обробляти довгі контекстні залежності, що робить його ефективним для реальних застосувань.

Трансформери – це архітектура нейронних мереж, яка використовує механізм самоуваги (self-attention) для обробки послідовностей даних. Вони вперше були представлені у моделях BERT (Bidirectional Encoder Representations from Transformers) і GPT (Generative Pre-trained Transformer) від Google.

Трансформери складаються з кількох шарів самоуваги і обробляють кожен елемент вхідної послідовності, враховуючи всі інші елементи одночасно. Це дозволяє моделі зберігати і використовувати довгострокові залежності. Механізм самоуваги дозволяє моделі динамічно визначати, які частини вхідної послідовності є найбільш важливими для поточного кроку обробки, що підвищує точність і ефективність розпізнавання. Моделі на основі трансформерів, такі як Wav2Vec від Facebook і Speech-Transformer, успішно застосовуються у задачах розпізнавання мовлення, забезпечуючи високу точність.

Convolutional Neural Networks (CNN) використовують згорткові шари для обробки даних і виділення важливих ознак. Вони часто застосовуються для попередньої обробки аудіосигналів і виділення корисних ознак. CNN складаються з послідовності згорткових шарів, пулінгових шарів і повнозв'язних шарів. Згорткові шари виділяють локальні ознаки, тоді як пулінгові шари зменшують розмірність даних. CNN ефективно виділяють важливі ознаки з вхідних аудіоданих, що покращує загальну продуктивність моделі.

Recurrent Neural Networks (RNN), включаючи варіанти LSTM та GRU (Gated Recurrent Units), здатні обробляти послідовності даних, зберігаючи інформацію про попередні етапи. RNN містять циклічні зв'язки, що дозволяє інформації циркулювати між нейронами. LSTM і GRU додають спеціальні механізми для контролю потоку інформації, що допомагає запобігти проблемам зникання градієнта. RNN можуть зберігати довгострокові залежності в послідовностях, що важливо для розпізнавання мовлення.

Енд-то-енд моделі дозволяють безпосередньо перетворювати аудіосигнал у текст без проміжних етапів виділення ознак або фонетичного моделювання. Енд-то-енд моделі часто використовують комбінацію CNN, RNN і трансформерів для обробки аудіосигналів і генерації тексту. Tacotron від Google – це приклад енд-то-енд моделі, яка використовується для синтезу мовлення, але аналогічні підходи застосовуються і для розпізнавання мовлення.

3.4.2 Переваги використання нейронних мереж.

Переваги використання нейронних мереж в STT

- Глибокі нейронні мережі здатні аналізувати складні мовні залежності, досягаючи високої точності навіть у випадках з шумом чи акцентами
- Навчання на великих обсягах даних дозволяє моделям постійно вдосконалювати свої результати і адаптуватися до нових мовних умов
- Моделі можуть навчатися на мультимовних даних і враховувати різні акценти та діалекти, роблячи їх універсальними для застосування в різних регіонах
- Використання RNN та трансформерів дозволяє моделям враховувати контекст мовлення, покращуючи точність в складних ситуаціях
- Моделі можуть розпізнавати мовлення навіть у присутності фонових шумів, що робить їх ефективними у реальних умовах
- Нейронні мережі можуть обробляти варіації в швидкості мовлення, інтонації та емоційному забарвленні голосу
- Паралельні обчислення і оптимізації дозволяють моделям працювати в режимі реального часу, що важливо для інтерактивних додатків
- Енд-то-енд підхід спрощує архітектуру системи, знижує затримки і підвищує ефективність обробки даних
- Нейронні мережі можуть інтегруватися з технологіями обробки природної мови, створюючи комплексні рішення для аналізу мовлення
- Моделі можуть навчатися новим задачам і адаптуватися до нових даних без значної перебудови системи

3.5 Оптимізація розпізнавання мови

3.5.1 Методи покращення точності та швидкості розпізнавання.

Оптимізація розпізнавання мови включає різні методи, спрямовані на покращення точності та швидкості системи розпізнавання мовлення (STT). Розглянемо основні методи, які застосовуються для досягнення цих цілей:

- Збір та анотація великих обсягів даних для покращення здатності моделей до генералізації
- Очистка даних шляхом видалення шумів, помилок і непотрібних даних з навчальних наборів
- Додавання шумів для симуляції різних умов запису і покращення стійкості моделей до шуму
- Зміна швидкості і тону мовлення в навчальних даних для покращення стійкості до варіацій у реальному мовленні
- Збільшення кількості шарів і нейронів у моделі для кращого захоплення складних залежностей у даних
- Використання гібридних архітектур, таких як поєднання CNN для виділення ознак і RNN або трансформерів для обробки послідовностей
- Використання методу Dropout для випадкового відключення нейронів під час навчання і запобігання перенавчанню моделі
- Застосування Batch normalization для нормалізації вхідних даних і стабілізації процесу навчання
- Використання адаптивних методів оптимізації, таких як Adam, RMSprop або AdaGrad, для швидшого і точнішого налаштування ваг моделі
- Поступове зменшення швидкості навчання (learning rate) для досягнення кращої конвергенції моделі
- Використання попередньо навчених моделей (наприклад, BERT, GPT) як базової архітектури з подальшим тонким налаштуванням
- Налаштування попередньо навчених моделей (файн-тюнінг) на нових даних для адаптації до специфічних умов
- Використання графічних процесорів (GPU) або тензорних процесорів (TPU) для паралельного навчання і інференції моделей
- Розподілення процесу навчання на кілька обчислювальних вузлів для швидшої і ефективнішої обробки великих обсягів даних

- Використання алгоритму beam search для декодування і знаходження найбільш ймовірних послідовностей слів
- Інтеграція мовних моделей (наприклад, N-грамових або трансформерних моделей) у процес декодування для врахування контексту і покращення якості розпізнаного тексту

3.5.2 Редукція шумів та покращення якості звуку

Редукція шумів і покращення якості звуку є критично важливими завданнями в системах розпізнавання мови (STT). Висока якість звукових даних значно покращує точність розпізнавання.

Попередня обробка звуку включає фільтрацію шумів та нормалізацію рівня звуку. Для видалення небажаних частот, які знаходяться поза діапазоном корисного сигналу, застосовуються фільтри нижніх і верхніх частот, а смугові фільтри пропускають лише частоти, що містять корисну інформацію. Нормалізація рівня звуку регулює амплітуду звукового сигналу, щоб уникнути впливу занадто тихих або занадто гучних звуків, забезпечуючи стабільний вхідний сигнал для моделі.

Методи редукції шумів включають спектральне віднімання, вейвлет-перетворення та адаптивні фільтри. Спектральне віднімання аналізує спектр шуму в тиші і віднімає його з основного сигналу під час мовлення, що ефективно для сталих шумів, таких як шум вентилятора або кондиціонера. Вейвлет-перетворення розділяє сигнал на вейвлети, виділяючи корисні компоненти мовлення і зменшуючи шуми, що добре підходить для виділення низькоінтенсивних мовних сигналів із шумного оточення. Адаптивні фільтри, такі як LMS (Least Mean Squares) і RLS (Recursive Least Squares), використовують адаптивні алгоритми для зменшення шумів у режимі реального часу, налаштовуючись під поточні умови шуму.

Аугментація даних для стійкості до шумів включає синтетичне додавання шумів і використання різних типів шумів. Додавання штучних шумів до навчальних даних допомагає тренувати модель для роботи у різних умовах.

Використання різноманітних шумів, таких як вуличний шум, шум натовпу та природні звуки, підвищує стійкість моделі до різних шумових умов.

Акустичні моделі, стійкі до шумів, використовують глибокі нейронні мережі, такі як автоенкодери, для автоматичного видалення шумів зі звукових сигналів. Архітектури нейронних мереж, такі як WaveNet і Wave-U-Net, спеціально розроблені для обробки та очищення звукових сигналів.

Покращення якості звуку включає методи upsampling і суперрезолюції, які підвищують роздільну здатність звукового сигналу, допомагаючи краще відновити мовні компоненти. Еквалізація регулює рівні різних частотних компонентів звукового сигналу для досягнення більш збалансованого і чистого звуку.

Використання багатоканальних записів передбачає застосування мікрофонних решіток, що дозволяє записувати звук за допомогою кількох мікрофонів. Це дає можливість застосовувати методи просторової фільтрації для відокремлення мовлення від шумів. Beamforming – це техніка обробки сигналу, яка використовує кілька мікрофонів для формування напрямку чутливості і придушення шумів з інших напрямків.

Постобробка звукових даних включає видалення ехоефектів і компресію динамічного діапазону. Використання методів цифрової обробки сигналів для зменшення ефектів ехо покращує чіткість мовлення. Компресія динамічного діапазону регулює різницю між тихими і гучними частинами звукового сигналу, роблячи його більш рівномірним для розпізнавання.

4. РЕАЛІЗАЦІЯ НАСКРІЗНОХ СИСТЕМИ STT ТА TSS

4.1 Апаратна конфігурація

Для дослідження та тестування системи розпізнавання мови (STT) та синтезу мови (TTS) було використано одноплатний комп'ютер Banana Pi M5, рис. 4.1. Основною операційною системою для стенду було обрано Debian 10, що забезпечило стабільне і гнучке середовище для розгортання необхідного програмного забезпечення.

Рисунок 4.1 – стенд для реалізації настрізної системи STT та TTS

Одноплатний комп'ютер Banana Pi M5 має наступні технічні характеристики:

- Процесор: Amlogic S905X3, чотириядерний ARM Cortex-A55
- Оперативна пам'ять: 4 ГБ DDR4
- Вбудована пам'ять: 16 ГБ eMMC
- Підтримка мікро-SD карт до 256 ГБ
- Порти: HDMI, USB 3.0, USB 2.0, Gigabit Ethernet, аудіовихід

На Banana Pi M5 було встановлено операційну систему Debian 10, яка забезпечує надійну платформу для розробки і тестування. Для написання тестових

програм і запуску моделей STT і TTS використовувалася мова програмування Python.

4.2 Тестування STT систем

Тестування системи STT включає кілька важливих кроків. Спочатку перевіряють, наскільки точно система перетворює аудіо в текст, порівнюючи результат із оригіналом. Далі тестують, як швидко система може перетворювати мовлення в текст. Код програми тестування наведений в Додаток А.

Обрані моделі для перевірки :

- Whisper small
- Whisper tiny
- wav2vec2-xls-r-300m-uk-with-wiki-lm
- wav2vec2-xls-r-base-uk-with-small-lm
- neon-stt-plugin-nemo

Рисунок 4.2 – діаграма проведення тестів

В таблицях 4.1, 4.2, 4.3 наведені результати перевірки роботи трьох популярних систем на українській мові.

Таблиця 4.1 – Результати тестування STT Whisper uk

N	Модель	Текст	Кодований текст	Точність, WER	Час обробки, сек
1	2	3	4	5	6
1	small	Завдяки	Завдяки!	1.0	52.85
2	small	Завдяки прогресивним та надійним продуктам та послугам, талановитим співробітникам і	завдяки прогресивним та надійним продуктам та послугам, талановитим співробітникам і серйозному підходу до	0.13	128.119

		серйозному підходу до глобального громадянства	глобального громадянства.		

Продовження Таблиці 4.1

N	Модель	Текст	Кодований текст	Точність, WER	Час обробки, сек
1	2	3	4	5	6
3	tiny	Завдяки	Завдяки.	1.0	7.38
4	tiny	Завдяки прогресивним та надійним продуктам та послугам, талановитим співробітникам і серйозному підходу до глобального громадянства	Завдяки прогрессивным танадийным продуктом Тапослухом. Талоновым спевробитником и серьезному петходу глобального громадянства.	0.86	18.67

Таблиця 4.2 – Результати тестування STT Wav2Vec2 uk

N	Модель	Текст	Кодований текст	Точність, WER	Час обробки, сек
1	2	3	4	5	6
1	Yehor/wav2vec2-xls-r-300m-uk-with-wiki-lm	Завдяки	Завдяки.	1.0	5.02
2	Yehor/wav2vec2-xls-r-300m-uk-with-wiki-lm	Завдяки прогресивним та надійним продуктам та послугам, талановитим співробітникам і серйозному підходу до глобального громадянства	завдяки прогресивним та надійним продуктам та послугам талановитим співробітникам і серйозному підходу до глобального громадянства	0.13	18.263
3	Yehor/wav2vec2-	Завдяки	завдяки	1.0	1.6

	xls-r-base-uk-with-small-lm				

Продовження Таблиці 4.2

N	Модель	Текст	Кодований текст	Точність, WER	Час обробки, сек
1	2	3	4	5	6
4	Yehor/wav2vec2-xls-r-base-uk-with-small-lm	Завдяки прогресивним та надійним продуктам та послугам, талановитим співробітникам і серйозному підходу до глобального громадянства	завдяки про гресивним та надійним продуктам та послугам талановитим зпівробітникам і серйозному підходу до глобального громадянства	0.33	6.80

В таблицях 4.1, 4.2, 4.3 видно як застосування різних наборів даних для обробки впливає безпосередньо на час обробки та точність результату.

Таблиця 4.3 - Результати тестування STT Neon Nemo

N	Модель	Текст	Кодований текст	Точність, WER	Час обробки, сек
1		Завдяки	завдяки	1.0	0.51
2		Завдяки прогресивним та надійним продуктам та послугам, талановитим співробітникам і серйозному підходу до глобального громадянства	завдяки прогресивним та надійним продуктам та послугам талановитим співробітникам і серйозному підходу до глобального громадянства	0.13	1.39

Рисунок 4.3 – Графік порівняння швидкостей обробки тексту на різних STT

Таблиця 4.4 – Результати тестування STT Whisper en

N	Модель	Текст	Кодований текст	Точність, WER	Час обробки, сек
1	small	variations	Variations	1.0	48.79
2	small	There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration	There are many variations of passages of Lormipsum available, but the majority have suffered alteration..	0.13	61.42
3	tiny	variations	variations.	1.0	6.63
4	tiny	There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration	There are many variations of passages of Laura Mipsom available, but the majority have suffered alteration.	0.18	9.325

Рисунок 4.4 – Порівняння короткого тексту whisper на різних мовах

Рисунок 4.5 – Порівняння короткого тексту whisper на різних мовах

Таблиця 4.5 – Результати тестування STT Wav2Vec2 en

N	Модель	Текст	Кодований текст	Точність, WER	Час обробки, сек
1	facebook/wav2vec2-base-960h	variations	VARIATIONS	1.0	1.68
2	facebook/wav2vec2-base-960h	There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration	THERE ARE MANY VARIATIONS OF PASSAGES OF LORAMIPSEM AVAILABLE BUT THE MAJORITY HAVE SUFFERED ALTERATION	0.12	5.39

Рисунок 4.6 – Порівняння короткого тексту whisper на різних мовах

В процесі тестування систем розпізнавання мови (STT) на платформі Banana Pi M5 були проведені детальні експерименти з використанням моделей Whisper (small, tiny), Wav2Vec 2.0 та NeMo STT. Для оцінки продуктивності моделей було

випробувано розпізнавання мовлення різної довжини, від коротких фраз до довгих абзаців. Це дозволило виявити здатність моделей до обробки як коротких, так і тривалих мовних сигналів. Крім того, моделі були протестовані на різних мовах, щоб визначити їхню швидкість і точність при різних налаштуваннях програми. Whisper, Wav2Vec 2.0 продемонстрували високу адаптивність до різних мов і акцентів, що підтвердило їхню універсальність і ефективність в умовах реального використання. Але час обробки на портативних пристроях занадто великий для того, що б їх використовувати. Тестування показало, що, незважаючи на обмежені апаратні ресурси Banana Pi M5, модель Neon STT забезпечує високу точність і швидкість розпізнавання мовлення, роблячи її придатною для використання в різних прикладних задачах.

4.3 Тестування TTS систем

В рамках дослідження системи синтезу мови (TTS) було проведено тестування моделі Neon на платформі Banana Pi M5. Основною метою тестування було оцінити швидкість роботи моделі при синтезі мовлення. Для цього було використано тексти різної довжини та складності, включаючи як короткі фрази, так і довгі абзаци. Результати тестування показали, що модель Neon забезпечує швидкий синтез мовлення, що дозволяє її використовувати в режимі реального часу. Незважаючи на обмежені апаратні ресурси Banana Pi M5, модель продемонструвала високу продуктивність, швидко обробляючи введені текстові дані і генеруючи відповідний аудіо вихід. Це підтверджує ефективність моделі Neon для задач синтезу мови в умовах реального використання.

Таблиця 4.6 – Результати тестування TTS Neon

N	Модель	Текст	Час обробки, сек
1		Завдяки	1.42
2		Завдяки прогресивним та надійним продуктам та послугам, талановитим співробітникам і	5.10

		серйозному підходу до глобального громадянства	

4.4 STT – TTS система

В рамках дослідження було проведено тестування системи, яка послідовно виконує розпізнавання мови (STT) і синтез мови (TTS) на платформі Banana Pi M5. Метою цього тестування було оцінити загальні затримки у часі, що виникають під час процесу перетворення мовлення у текст і назад у мовлення. Для цього використовувалися модель Neon для TTS та STT як найшвидші для української мови. Щоб пришвидшити обробку, в системі було використано технологію виявлення голосової активності (Voice Activity Detection, VAD), яка допомагає ідентифікувати моменти мовлення та відсікати періоди тиші, тим самим зменшуючи обсяг оброблюваних даних

Рисунок 4.7 – блок-схема для перевірки затримки

Таблиця 4.7 – результати тестування

N	Слово	Довжина слова	STT, сек	TTS, сек	Затримка сумарна, сек
1	Завдяки	7	0.48	2.96	3,44
2	університет	11	0.53	4.14	4,67
3	талановитим	11	0.49	2.79	3,28
4	так	3	0.41	3.39	3,8

Тестування показало, що використання VAD суттєво знижує затримки в обробці, особливо для довгих записів з великими періодами тиші. Моделі STT і

TTS демонстрували високу продуктивність, обробляючи текстові дані швидко і з мінімальними затримками.

Також було знайдено слова, які неправильно розпізнаються stt системою. Для вирішення цього питання потрібно навчати модель на більших наборах даних. До прикладу, слова іноземного походження “трамвай” та “тролейбус”.

Рисунок 4.7 – неправильно розпізнані слова

5. РОЗДІЛ СТАРТАП-ПРОЕКТУ

5.1 Вступ

Метою даного проекту є розробка та впровадження системи, що використовує мобільний ПК як голосовий кодек для GSM з функціональностями TTS та STT. Така система може бути використана в різних сферах, включаючи телекомунікації, розумні домашні системи, підтримку клієнтів тощо.

Сильні сторони:

- Надзвичайно гнучка система завдяки відкритій архітектурі апаратного та програмного забезпечення. Це дозволяє легко адаптувати систему під конкретні потреби користувачів, інтегрувати нові функції та модулі. Наприклад, можна додавати нові алгоритми для розпізнавання мови чи синтезу мови, змінювати налаштування для оптимізації продуктивності, або інтегрувати з іншими системами для розширення функціональності.
- Економічно вигідніший варіант у порівнянні з іншими одноплатними комп'ютерами, такими як Raspberry Pi. Це знижує загальну вартість системи, роблячи її доступнішою для широкого кола користувачів, включаючи невеликі підприємства та освітні заклади. Використання відкритих програмних продуктів також знижує витрати на ліцензії.
- Система дозволяє користувачам налаштовувати її під свої специфічні потреби, що є великим плюсом для підприємств та приватних осіб, які мають унікальні вимоги. Наприклад, можна налаштувати специфічні голосові команди для управління домашніми пристроями або налаштувати систему для роботи з різними мовами та діалектами.

Слабкі сторони:

- Система є складною та гнучкою, вона потребує постійної технічної підтримки для забезпечення стабільної роботи. Це включає оновлення програмного забезпечення, усунення технічних несправностей, налаштування нових функцій та забезпечення безпеки системи.

- Залежність від якості та стабільності GSM мережі для передачі даних та голосових повідомлень. У випадку поганого покриття або перебоїв у роботі мережі, якість зв'язку може значно знизитись, що впливає на загальну продуктивність системи. Це особливо важливо в регіонах з нестабільною мережею або відсутністю надійного покриття.

Можливості:

- Проект має потенціал для виходу на міжнародний ринок завдяки своїй гнучкості, економічній доступності та універсальності. Інтеграція з різними мовами та культурами дозволить залучити ширшу аудиторію, включаючи користувачів з різних країн. Це може бути здійснено через міжнародні партнерства, участь у виставках та конференціях, а також через онлайн-продажі.
- Проект має можливість інтеграції з новітніми технологіями, такими як штучний інтелект, Інтернет речей (IoT) та хмарні обчислення. Це дозволить розширити функціональність системи, підвищити її ефективність та забезпечити нові можливості для користувачів. Наприклад, використання AI для поліпшення розпізнавання мови або аналізу даних може значно підвищити продуктивність системи.

Загрози:

- Великі технологічні компанії, такі як Google, Amazon та Microsoft, мають значні ресурси та можливості для розробки і впровадження подібних технологій. Вони можуть запропонувати більш інтегровані та масштабовані рішення, що може створити конкуренцію для вашої системи. Це вимагає постійного вдосконалення продукту та пошуку нових унікальних переваг для збереження конкурентоспроможності.
- Зміни у законодавстві та регуляторному середовищі можуть вплинути на роботу системи. Наприклад, нові вимоги до безпеки даних або обмеження на використання певних технологій можуть вимагати додаткових ресурсів

для відповідності новим нормам. Це може створити додаткові витрати та затримки у впровадженні системи.

Зміст ідеї

Таблиця 5.1 – Опис ідеї проекту

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Захищений зв’язок	Забезпечення безпечного зв'язку для військових операцій та державних службовців, щоб уникнути витоку конфіденційної інформації. Захист корпоративних комунікацій, що дозволяє запобігти промисловому шпіонажу та захистити комерційні таємниці. Використання приватними особами для забезпечення захисту особистих розмов та даних, особливо для журналістів, активістів та інших осіб, що можуть бути під загрозою.	Забезпечення високого рівня захисту даних, що передаються, що мінімізує ризики перехоплення та несанкціонованого доступу. Гарантія конфіденційності комунікацій, що важливо для бізнесу, уряду та приватних осіб. Рішення, що може бути інтегроване в існуючі системи зв'язку без значних витрат на модернізацію обладнання.

5.2 Кошторис

Таблиця 5.2 – орієнтовний кошторий проекту

Компонент	К-ть	Ціна	Загальна ціна
Banana pi m5	1	5200 грн	5200 грн
GSM	1	1000 грн	1000 грн
3D Sound USB	1	70 грн	70 грн
Навушники	1	160 грн	160 грн

Акумулятор/powerbank	1	600 грн	600 грн
Кабель type c для акумулятора	1	200 грн	200 грн

Порівняння з аналогами

Таблиця 5.3 – порівняння з аналогами

N	Технічно-економічні характеристики ідеї	Пристрій	Конкурент 1	Конкурент 2	Конкурент 3	Слабкість	Нетраль	Сила
1	Ціна	180$	200$	500$	50$		+	
2	Призначення	Захищений зв’язок	Мобільний зв’язок	Зв’язок на далекі відстані у дикій природі	Обробка сигналів			+
3	Портативність	Простий для транспортування	Простий для транспортування	Простий для транспортування	Простий для транспортування		+	
4	Технологічність	Запуск системи 1 кнопкою	Запуск системи 1 кнопкою	Запуск системи 1 кнопкою	Обробка сигналу проводиться окремо		+	
5	Безпечність	Безпечна	Невідомо	Безпечна	Безпечна		+	

Конкурент 1 – мобільний телефон

Конкурент 2 – портативна рація

Конкурент 3 – sdr

5.3 Технологічний аудит ідеї проекту

Таблиця 5.4 – аудит технологій для проекту

N	Ідея проекту	Технології	Наявність технологій	Доступність технологій
1	Захищений зв’язок	GSM зв’язок	Усі компоненти розроблені	Компоненти пристрою обираються з наявності в магазині та відповідно до технічних характеристик.
		Шифрування	Усі компоненти розроблені	Алгоритми розроблені. Потрібно вибирати з актуальних на сьогодні.
		STT	Усі компоненти розроблені	Алгоритми розроблені. Потрібно вибирати з актуальних на сьогодні. Також багато алгоритмів можливо адаптувати.
		TTS	Усі компоненти розроблені	Алгоритми розроблені. Потрібно вибирати з актуальних на сьогодні. Також багато алгоритмів можливо адаптувати.
		Портативний ПК	Усі компоненти розроблені	Компоненти пристрою обираються з наявності в магазині та відповідно до технічних

характеристик.

5.4 Аналіз ринкових можливостей

Таблиця 5.5 – таблиця ринкових потреб та попиту

N	Потреба, що формує ринок	Цільові сегменти ринку	Відмінності у поведінці ЦА	Вимоги споживачів до товару
1	Популярність домашніх асистентів	Розробники розумних домашніх систем: Використовують технології для управління побутовими приладами, безпеки та комфорту.	Розробники розумних домашніх систем: Зосереджуються на інтеграції з іншими IoT пристроями та простоті використання для кінцевих споживачів.	Розробники розумних домашніх систем: Простота інтеграції, зручний інтерфейс для кінцевих користувачів, безпека.
2	Оптимізація бізнесу	Бізнес та служби підтримки: Потребують інструментів для автоматизації обслуговування клієнтів та підвищення ефективності обробки запитів.	Бізнес та служби підтримки: Віддають перевагу рішенням, які дозволяють знизити витрати на персонал та підвищити швидкість обслуговування клієнтів.	Бізнес та служби підтримки: Ефективність автоматизації, зниження операційних витрат, швидка реакція на запити клієнтів.
3	Інтерактивність навчального процесу	Освітні установи: Використовують для дистанційного навчання та підтримки студентів з обмеженими можливостями	Освітні установи: Орієнтуються на адаптивність технології до навчальних програм і потребують високого рівня персоналізації	Освітні установи: Адаптивність до різних навчальних програм, підтримка кількох мов, можливість інтеграції з

			для підтримки різних категорій студентів.	іншими освітніми платформами.

5.5 Висновки

Проект, що використовує Banana Pi для створення системи захищеного зв'язку з функціями TTS і STT, має значний комерційний потенціал. Існує стабільний попит на технології захищеного зв'язку серед військових, урядових установ, бізнесу та приватних осіб. Ринок технологій захищеного зв'язку зростає, що створює сприятливі умови для впровадження проекту.

Основні переваги системи включають економічну доступність, гнучкість та можливість індивідуального налаштування, що дозволяє конкурувати з великими гравцями на ринку. Бар’єри входження пов’язані з необхідністю забезпечення високого рівня безпеки та відповідності регуляторним вимогам.

Враховуючи високий попит та динаміку ринку, проект має перспективи для успішної ринкової реалізації та подальшої імплементації, що дозволить задовольнити потреби різних сегментів ринку та підвищити ефективність їх діяльності.

ЗАГАЛЬНІ ВИСНОВКИ

У даній роботі було розглянуто процеси перетворення сигналів у голосовому каналі системи GSM, зокрема параметричне перетворення мовленнєвих сигналів у вокодерних системах та системах speech-to-text (STT) і text-to-speech (TTS).

Було проведено аналіз існуючих методів модуляції даних для їх передачі через голосові канали системи GSM та запропоновано ряд рішень щодо їх оптимізації.

Розроблено концепцію побудови системи передачі шифрованого мовлення через голосові канали GSM з використанням систем STT та TSS на її вході і виході. Основні риси розробленої концепції такі:

- Мінімізація часу обробки системи STT для побудови каналу зв'язку в рельаному часі
- Мінімізація часу обробки системи TTS для побудови каналу зв'язку в рельаному часі

Було запропоновано алгоритм розпізнавання та синтезу мовлення в режимі реального часу, який дозволяє проводити розпізнавання та синтез мовлення як в режимі розпізнавання окремих файлів та аудіо. Так і працювати в реальному часі з голосомє

На підставі багатьох звітів стосовно експлуатації тестових зразків, проведено аналіз роботи систем та висунуто основні вимоги до роботи нових версій ПЗ.

Було обрано необхідні інструменти і технології для розробки ПЗ, створено оновлений каркас ПЗ та проведені зміни у конфігураціях системи для забезпечення стабільної роботи.

Реалізовано алгоритм мережевої взаємодії з серверною ЕОМ на стороні клієнта та налаштовано розроблене ПЗ і середовище розробки для впровадження алгоритму розпізнавання та синтезу мовлення.

Реалізовано алгоритми навчання та розпізнавання мовлення з урахуванням особливостей платформи Android. Проведено початкове налаштування і підготовку середовища виконання ПЗ до подальшої роботи.

Таким чином, проведене дослідження підтвердило можливість та ефективність застосування систем STT та TSS для захищеного голосового зв'язку через голосові канали GSM, що відкриває нові перспективи для розвитку захищених комунікаційних технологій на базі портативних мікроелектронних пристроїв.

ПЕРЕЛІК ПОСИЛАНЬ

1. Klatt D. H. Review of text-to-speech conversion for English / Dennis H. Klatt., 1987.
2. Methods, Techniques, and Algorithms [Електронний ресурс] – Режим доступу до ресурсу: http://research.spa.aalto.fi/publications/theses/lemmetty_mst/chap5.html.
3. Nilashree W. S. Speech Communication: Human and Machine / W. S. Nilashree, S. S. Milind., 2014. – (Open Access Library Journal). – (1; 4).
4. Laine U. PARCAS, A new terminal analog model for speech synthesis / U. Laine. – Tampere, Finland, 1982.
5. Diemo S. Concatenative Sound Synthesis: The Early Years / Schwarz Diemo., 2006. – (Journal of New Music Research). – (35; № 1).
6. Pitch-Synchoronous Overlap-Add (PSOLA) [Електронний ресурс] // speechprocessingbook – Режим доступу до ресурсу: https://speechprocessingbook.aalto.fi/Representations/Pitch-Synchoronous_Overlap-Add_PSOLA.html.
7. Statistical parametric speech synthesis [Електронний ресурс] // speechprocessingbook – Режим доступу до ресурсу: https://speechprocessingbook.aalto.fi/Synthesis/Statistical_parametric_speech_synthesis.html.
8. Karen S. WAVENET: A GENERATIVE MODEL FOR RAW AUDIO / S. Karen, G. Alex.
9. What Is GSM [Електронний ресурс] // spiceworks. – 2022. – Режим доступу до ресурсу: https://www.spiceworks.com/tech/networking/articles/what-is-gsm/.
10. Barry D. L. MODEM FOR COMMUNICATING DATA / Dorr L. Barry., 2007.

ДОДАТОК А

python програма для оцінки швидкості та точності stt whisper

```
import time
import json
import numpy as np
import whisper

audio_samples = ["audio1.wav", "audio2.wav"]
transcripts = ["transcript1.txt", "transcript2.txt"]

model = whisper.load_model("tiny")

def load_transcript(file_path):
    with open(file_path, "r") as file:
        return file.read().strip()

def calculate_wer(reference, hypothesis):
    ref_words = reference.split()
    hyp_words = hypothesis.split()
    d = np.zeros((len(ref_words) + 1, len(hyp_words) + 1))
    for i in range(len(ref_words) + 1):
        d[i][0] = i
    for j in range(len(hyp_words) + 1):
        d[0][j] = j
    for i in range(1, len(ref_words) + 1):
        for j in range(1, len(hyp_words) + 1):
            if ref_words[i - 1] == hyp_words[j - 1]:
                d[i][j] = d[i - 1][j - 1]
            else:
                d[i][j] = min(d[i - 1][j] + 1, d[i][j - 1] + 1, d[i - 1][j - 1] + 1)
    return d[len(ref_words)][len(hyp_words)] / len(ref_words)
```

```
def transcribe_audio(file_path):
    transcription = model.transcribe(file_path)
    return transcription["text"].strip()

def measure_latency(audio_file):
    start_time = time.time()
    transcribe_audio(audio_file)
    return time.time() - start_time

def run_tests():
    results = {
        "Точність": [],
        "Затримка": [],
        "Надійність": [],
        "Слова": []
    }

    for i, audio_file in enumerate(audio_samples):
        print(f"Testing {audio_file}...")
        reference = load_transcript(transcripts[i])
        transcription = transcribe_audio(audio_file)
        wer = calculate_wer(reference, transcription)
        latency = measure_latency(audio_file)

        results["Точність"].append(wer)
        results["Затримка"].append(latency)
        results["Надійність"].append(transcription)
        results["Слова"].append(len(transcription.split()))

    return results

if __name__ == "__main__":
    test_results = run_tests()
```

```
print(json.dumps(test_results, indent=4, ensure_ascii=False))
```

ДОДАТОК Б

```
# python програма для оцінки швидкості TTS neon
import time
from neon_tts_plugin_coqui import CoquiTTS
import IPython

tts = CoquiTTS("uk")

start_time = time.time()
ipython_dict = tts.get_audio("Завдяки", audio_format="ipython")
stop_time = time.time()
delay = stop_time - start_time
print(f"Обробка TTS: {delay} c")
```

ДОДАТОК В

```
# python програма для оцінки швидкості системи STT -> TTS
import pyaudio
import webrtcvad
import numpy as np
import threading
from neon_stt_plugin_nemo import NemoSTT
import speech_recognition as sr
from neon_tts_plugin_coqui import CoquiTTS
import IPython
import asyncio
import queue
import time

stt = NemoSTT()
tts = CoquiTTS("uk")
audio_queue = queue.Queue()
text_queue = queue.Queue()


FORMAT = pyaudio.paInt16
CHANNELS = 1
RATE = 16000
CHUNK = 320

p = pyaudio.PyAudio()
p_tts = pyaudio.PyAudio()
stream_tts = p_tts.open(format=FORMAT,
                channels=CHANNELS,
                rate=RATE,
                output=True,
                frames_per_buffer=CHUNK)
vad = webrtcvad.Vad()
```

```
vad.set_mode(3)

def global_tts():
    while 1:
        text = text_queue.get()
        start_time = time.time()
        ipython_dict = tts.get_audio(text, audio_format="ipython")
        audio_data = ipython_dict["data"]
        sample_rate = ipython_dict["rate"]
        data_int16 = (np.array(audio_data) * (2**15 - 1)).astype(np.int16)
        stream_tts.write(data_int16.tobytes())
        stop_time = time.time()
        delay = stop_time - start_time
        print(f"Обробка TTS: {delay} c")


def global_stt():
    while 1:
        bufff = audio_queue.get()
        start_time = time.time()
        audio_data = sr.AudioData(np.int16(bufff).tobytes(), sample_rate=RATE, sample_width=2)
        text = stt.execute(audio_data, 'uk')
        if len(text) > 0:
            stop_time = time.time()
            delay = stop_time - start_time
            print(f"Обробка STT: {delay} c")
            text_queue.put(text)



    print(text)
```

```
def record_audio():
    stream = p.open(format=FORMAT,
                    channels=CHANNELS,
                    rate=RATE,
                    input=True,
                    frames_per_buffer=CHUNK)

    print("Recording...")

    try:
        bufff = []
        while 1:
            data = stream.read(CHUNK)
            audio_array = np.frombuffer(data, dtype=np.int16)
            if vad.is_speech(audio_array.tobytes(), sample_rate=RATE):
                bufff.append(audio_array)
            else:
                if len(bufff):
                    audio_queue.put(bufff)
                    bufff = []
    except KeyboardInterrupt:
        pass

    p.terminate()


record_thread = threading.Thread(target=record_audio)
stt_thread = threading.Thread(target=global_stt)
tts_thread = threading.Thread(target=global_tts)

record_thread.start()
stt_thread.start()
tts_thread.start()
```

ДОДАТОК Г

Інструкція для запуску Додаток А

1 – створення локального середовища розробки python venv p назвою local

python3.7 –venv local

2 – активація середовища

source local/bin/activate

3 – встановлення бібліотек для роботи через pip

3 – запуск програми. Огляд результатів

Бачимо список, що відповідає кількості файлів тестів. Присутній результат по точності, затримці обробки, кількості слів. Також виводяться кодований текст, для візуальної оцінки відповідності критерію похибки.

Внесення модифікацій в Додаток А дозволяє провести тестування для іншої STT моделі.

pi@bananapi: /mnt/sdcard/wav
(myenv) pi@bananapi:/mnt/sdcard/wav$ ls
audio1.wav audio3.wav main.py transcript1.txt transcript3.txt
audio2.wav audio4.wav myenv transcript2.txt transcript4.txt
(myenv) pi@bananapi:/mnt/sdcard/wav$ python main.py
Some weights of Wav2Vec2ForCTC were not initialized from the model checkpoint
ly initialized: ['wav2vec2.masked_spec_embed']
You should probably TRAIN this model on a down-stream task to be able to use it
Testing audio3.wav...
Testing audio4.wav...
{
"Точність": [
1.0,
1.0
],
"Затримка": [
1.6853070259094238,
5.395263433456421
],
"Надійність": [
"VARIATIONS",
"THERE ARE MANY VARIATIONS OF PASSAGES OF LORAMIPSEM AVAILABLE BUT THE
],
"Масштабованість": [
1,
15
]
}
(myenv) pi@bananapi:/mnt/sdcard/wav$

ДОДАТОК Д

Інструкція для запуску Додаток Б

1 – створення локального середовища розробки python venv p назвою local

python3.7 –venv local

2 – активація середовища

source local/bin/activate

3 – встановлення необхідних бібліотек для роботи через команду pip

4 – запуск програми. Огляд результатів

Відображається системна інофрмація про час обробки. І реальна, яка знімається програмою python.

ДОДАТОК Е

Інструкція для запуску Додаток В

1 – створення локального середовища розробки python venv р назвою local

python3.7 –venv local

2 – активація середовища

source local/bin/activate

3 – встановлення необхідних бібліотек для роботи через команду pip. Складову однакові для STT та TTS. Тому що в обох випадках йде робота й нейронними мережами.

4- після завантаження програми, з’явиться напис, який сповіщає про зчитування мікрофону