УДК 004.934

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет *Комп'ютерних наук*
(повна назва)

Кафедра *Системотехніки*
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

Рівень вищої освіти *другий (магістерський)*

Розробка методу аутентифікації користувача через голосове повідомлення
(тема)

Виконав:
Студент 2 курсу, групи *ІТПм-19-1*
Спеціальність *122 – Комп'ютерні науки*
(код і повна назва напряму)
Тип програми *освітньо-професійна*
(освітньо-професійна або освітньо-наукова)
Освітня програма *Інформаційні технології проектування*
(повна назва освітньої програми)

Зіміна А.Р.
(прізвище, ініціали)

Керівник доц. Губаренко Є.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ *Гребеннік І. В.*
(підпис) (прізвище, ініціали)

2020 р.

Харківський національний університет радіоелектроніки

Факультет *Комп'ютерних наук*

(повна назва)

Кафедра *Системотехніки*

(повна назва)

Рівень вищої освіти *другий (магістерський)*

Спеціальність *122 – Комп'ютерні науки*

(код і повна назва)

Тип програми *освітньо-професійна*

(освітньо-професійна або освітньо-наукова)

Освітня програма *Системне проектування*

(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____

(підпис)

«____»_____ 20___ р.

ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ

студентові *Зіміній Анні Романівні*

(прізвище, ім'я, по батькові)

1. Тема роботи *«Розробка програмного засобу розпізнавання голосу і голосових команд»*

затверджена наказом по університету від « *02* » *11* 2020 р. № *1517 Ст*

2. Термін подання студентом роботи (проекту) *21 грудня 2020 р.*

3. Вихідні дані до роботи (проекту) *Функція: Розробка програмного засобу розпізнавання голосу і голосових команд. Організація даних: файлова з прямим доступом . Форма діалогу: консольний додаток. Перелік використовуваних програмних засобів: ОС Microsoft Windows 7 та вище, інтегроване середовище програмування. Технічне забезпечення: комп'ютер з не менш, ніж 1Гб оперативної пам'яті*

4. Зміст пояснювальної записки (перелік питань, що потрібно розробити) *Вступ. Аналіз предметної області та постановка задачі. Опис та аналіз підходів до розробки методів ідентифікації голосу. Проектування і розробка програмного засобу. Аналіз результатів, отриманих за допомогою тестового програмного засобу. Висновки*

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслеників, плакатів) *Плакати на аркушах формату А4: класифікація систем розпізнавання мови, загальний алгоритм розпізнавання голосових, мовні технології, класифікація шумів у мовних сигналах, види клацань, алгоритм виявлення клацань, приклад перевантаження, контекстна діаграма, декомпозиція основних процесів, DFD діаграма, діаграма варіантів використання системи, діаграма класів, підгрузка словника та бібліотеки Sphinx4, демонстрація розпізнавання мови на тестових даних*

6. Консультанти розділів роботи (проекту)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	*Отримання завдання на атестаційне проектування*	*04.09.2020*	
2	*Аналіз завдання, пошук літератури та аналогів з теми*	*16.09.2020*	
3	*Опрацювання літератури та аналіз об'єкту дослідження*	*23. 10.2020*	
4	*Вибір оптимального підходу до розробки програмних засобів для автоматизованого тестування*	*26. 10. 2020*	
5	*Проектування програмного засобу*	*04.11. 2020*	
6	*Вибір програмних, мовних та технічних засобів*	*06.11. 2020*	
7	*Розробка програмного засобу*	*07.11. 2020*	
8	*Аналіз результатів, отриманих за допомогою програмного засобу*	*11.06. 2020*	
9	*Оформлення пояснювальної записки та програмної документації*	*13. 12. 2020*	
10	*Оформлення графічної частини та презентаційних матеріалів комп'ютерного захисту*	*14. 12. 2020*	
11	*Представлення на рецензування*	*15. 12. 2020*	

Дата видачі завдання *04 вересня 2020 р.*

Студент _____ *Зіміна А.Р.*
(підпис)

Керівник роботи _____ *доц. Губаренко Є.В*
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Атестаційна робота містить: 67 сторінки, 26 рисунів, 2 таблиці, 4 додатки, 27 джерел. Графічна частина атестаційної роботи містить 26 плакатів.

АЛГОРИТМ DTW, АУТЕНТИФІКАЦІЯ, БІОМЕТРІЯ, МОВНА МОДЕЛЬ, РОЗПІЗНАВАННЯ ГОЛОСУ, ФІЛЬТРАЦІЯ ШУМУ, SIMULINMK.

Об’єкт розробки – метод аутентифікаціЇ користувача через голосове повідомлення.

Мета атестаційної роботи – розробка методу аутентифікаціЇ користувача через голосове повідомлення.

Методи розробки – язик програмування Java, Java утиліта Simulink.

Результати атестаційної роботи – метод аутентифікаціЇ користувача через голосове повідомлення.

Область застосування – метод може бути використаний для авторизації на персональних пристроях за допомогою голосу.

ABSTRACT

Thesis contains: 67 pages, 26 figures, 2 tables, 4 applications, 27 sources. Graphic part of the thesis contains 26 posters.

AUTHENTICATION, BIOMETRY, DTW ALGORITHM, LANGUAGE MODEL, NOISE FILTRATION, VOICE RECOGNITION.

The object of development is a method of user authentication via voice message.

The purpose of certification work is to develop a method of user authentication via voice message.

Development methods – Java programming language, Java utility Simulink.

The results of certification work – a method of user authentication via voice message.

Scope – the method can be used for authorization on personal devices by voice.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ФДС – методом фонетичного декодування слів.

ВЧ-сигнали – високочастнотні сигнали.

ARM – мікропроцесорна архітектура.

VoiceXML – інтерактивна мова розмітки.

Ггц – гігагерц.

Гб – гігабайт.

ВСТУП

Найважливішим засобом комунікації є мова. Цей процес в свою чергу має на увазі формування, сприйняття і розуміння деяких мовних конструкцій.

У світі інформаційних технологій голосове управління породжує новий спосіб взаємодії з функціями різних пристроїв. Така можливість зручна в деяких ситуаціях, наприклад, коли не хочеться шукати інформацію в смартфоні або якщо необхідно звернутися до деякої функції пристрою на відстані (якщо, наприклад, потрібно зробити фотографію). Для створення такої технології необхідно було вирішити таке завдання, як розпізнавання мови.

В останні роки для ідентифікації особи людини все більш широке застосування отримують біометричні технології. Вони використовуються в системах розмежування доступу, при проведенні фінансових транзакцій, при запитах конфіденційної інформації по телефону, при управлінні різними пристроями, в криміналістиці і т.д. Використання біометричних технологій в цих сферах має низку істотних переваг перед традиційними засобами ідентифікації (наприклад, використанням пароля). В першу чергу до таких переваг належать висока надійність ідентифікації та зручність використання для людини.

Як ідентифікаційні параметри в біометричних технологіях використовуються фізіологічні та поведінкові характеристики людини. До таких характеристик відносяться відбитки пальців, голос, райдужна оболонка ока, обличчя людини, почерк і ін.

В даний час найбільш поширеними біометричними характеристиками людини є відбитки пальців і райдужна оболонка ока. У той же час голос використовується не так широко, хоча він має ряд істотних переваг, наприклад, простота зняття біометричного параметра (досить лише стандартного мікрофона), а також зручність використання.

На сьогоднішній день в світі існує ряд компаній, що займаються розробкою систем ідентифікації голосу. Досягнуті певні успіхи в цій галузі (ймовірність помилки ідентифікації 1 - 3%). Однак існуючі розробки мають ряд недоліків.

Алгоритми досить складні і вимагають великих обчислювальних ресурсів, що обмежує область їх застосування тільки високопродуктивними ЕОМ (час ідентифікації 3-5 секунд при тривалості фрази 3 секунди на ЕОМ з частотою процесора 1,60 ГГц і об'ємом оперативної пам'яті 896 МБ).

Крім того, у всіх системах відсутня можливість настройки алгоритмів під різні умови застосування (рівень шуму, особливість голосу конкретної людини, поріг помилок і т.д.).

Також, жоден розробник не надає кошти для тестування розробленої ним системи ідентифікації голосу, тоді як особливості умов застосування можуть значно впливати на якість роботи алгоритму.

Більшість алгоритмів не враховують текстовий зміст усного фрази (фонемную складову), виділяючи лише індивідуальні характеристики голосу, що значно знижує надійність ідентифікації.

З урахуванням сказаного попереду стоїть завдання розробки нової моделі ідентифікації голосового повідомлення по фонемній складовій і індивідуальних характеристиках голосу, вільної від зазначених недоліків, а також комплексу програм, що реалізує дану модель і дозволяє її тестувати.

Метою цього проекту є розробка математичної моделі ідентифікації голосового повідомлення по фонемной складовій і індивідуальних характеристик голосу, а також розробка комплексу програм, що реалізує дану модель. Цільова аудиторія користувачів такого програмного засобу – системи розмежування доступу, при проведенні фінансових транзакцій, при запитах конфіденційної інформації по телефону, при управлінні різними пристроями, в криміналістиці і т.д.

Виходячи з поставлених цілей, в роботі вирішуються наступні завдання:

– аналіз математичних методів, які можна застосувати для вирішення завдання ідентифікації голосового повідомлення;

– розробка математичної моделі ідентифікації голосового повідомлення по фонемной складової і індивідуальних характеристик голосу;

– програмна реалізація розробленої моделі ідентифікації голосового повідомлення;

– оцінка впливу значень варійованих параметрів (параметри моделі, за допомогою яких проводиться її настройка) розробленої моделі на якість ідентифікації;

– оцінка впливу різних вимовлених фраз на якість ідентифікації.

Методи дослідження запозичені з наступних областей:

– цифрова обробка сигналів;

– коливання і хвилі;

– математичний аналіз;

– математичне моделювання;

– чисельні методи;

– теорія ймовірностей і математична статистика;

– теорія мов програмування;

– теорія побудови баз даних.

Наукову новизну атестаційної роботи складають результати, отримані в ході вирішення поставлених завдань:

– модель ідентифікації голосового повідомлення по фонемній складовій і індивідуальних характеристиках голосу;

– метод поділу голосового повідомлення на фонеми;

– метод обробки фонем для їх порівняння;

– метод матричного аналізу порівняння фонем голосових повідомлень;

Розроблений метод може бути використаний для ідентифікаціїї голосу у різніх сферах.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ І ПОСТАНОВКА ЗАДАЧІ

1.1 Проблема розпізнавання вербальної інформації

Розпізнавання мови – процес перетворення мовного сигналу в цифрову інформацію (наприклад, текстові дані).

У будь-якої людини є свої особливі вокальні характеристики, які визначаються індивідуальною структурою його голосового апарату. Прислухаючись до розмови, людина може на рівні підсвідомості ідентифікувати голоси будь-яких інших людей, однак розробка автоматичного відрізнителя мови пов'язана зі значними труднощами.

Задача розпізнавання людини по голосу полягає у виділенні з вхідного аудиопотоку людської мови, її класифікацію і розпізнавання. При цьому зазвичай вирішуються дві підзадачі: розпізнавання мовця і перевірка. Для вирішення цих підзадач можна визначити метод розрахунку ступеня подібності вибірки з опорними сигналами. Ступінь подібності опорної і тестової вибірок можна розрахувати з використанням певної міри відстані або з використанням імовірнісних критеріїв [2]. Алгоритм ідентифікації мовця можна також визначити як текстозалежний і текстонезалежний. Якщо алгоритм ідентифікації мови залежить від тексту, то в ньому можна використовувати як фіксовані заздалегідь фрази, так і фрази, які генеруються системою розпізнавання. Текстонезалежні системи необхідні для обробки довільної мови.

Починаючи з 1980-х років в розпізнаванні мови найгостріше стоїть проблема наявності перешкод. Системи ефективно працюють в ідеальних умовах запису, але при цьому не справляються з фоновими шумами на кшталт звуку з сусідньої кімнати. Штучні шуми виділити цілком можливо, але важко відрізнити голос людини, який нам потрібно розпізнати, від голосу людини, що розмовляє по сусідству. Проблема завадостійкості до цього часу не вирішенах[3].

При створенні системи розпізнавання мови потрібно вибрати, який рівень абстракції адекватний поставленій задачі, які параметри звукової хвилі будуть

використовуватися для розпізнавання і методи розпізнавання цих параметрів. Розглянемо основні відмінності в структурі і процесі роботи різних систем розпізнавання мови.

У більшості існуючих механізмів можна виділити чотири основні модулі:

– модуль збору даних – включає отримання вхідного сигналу і його попередню обробку, яка може включати автоматичне регулювання посилення, виявлення присутності / відсутності мови і виявлення інтонаційного кінця фрази. Цей модуль включає також виділення відрізка мови з вхідного мовного сигналу.

– екстрактор – виконує частотний аналіз сигналу. Акустично-фонетічесій потік даних розбивається на короткі кадри, чи вектори, тривалістю близько 10 мс. Як правило, для кожного кадру визначається низка параметрів, якы використовують швидке перетворення Фур'є. Крім того, багато систем використовують замість або разом з цими характеристиками інші, наприклад, спектральні характеристики, а також першу і другу похідну від спектральних характеристик [4].

– компаратор – здійснює акустичне порівняння: кожен кадр, або вектор, порівнюється з наявними акустично-фонетичними зразкам, що зберігаються в спеціальній базі даних. При цьому порівнюватися можуть як окремі фонеми, так і слова, і навіть фрази. При невеликій кількості слів, використовуваних диктором, більш високу надійність і швидкість можна очікувати від розпізнавання цілих слів, але при збільшенні словника швидкість різко падає, і оптимальним стає розпізнавання окремих фонем.

– інтерпретатор вирішує завдання динамічного програмування з метою знайти найкращу розбивку отриманого від компаратора "алфавітного" потоку на слова і фрази. Залежно від обсягу використовуваного словника і діючих синтаксичних правил, застосовуються різні стратегії пошуку та відсіву.

Системи розпізнавання мови можна класифікувати в залежності від:

– призначення (системи диктування, командні системи);

– типу мови (злита або роздільна мова);

– розміру словника (обмежений набір слів, словник великого розміру);

– диктора (дікторозалежні і дикторонезалежної системи);

– механізму функціонування (найпростіші (кореляційні) детектори, експертні системи з різним способом формування і обробки бази знань, ймовірносно-мережеві моделі прийняття рішення, в тому числі нейронні мережі);

– використовуваного алгоритму (нейронні мережі, приховані Марковские моделі, динамічне програмування);

– типу структурної одиниці (фрази, слова, фонеми, діфони, Алофон);

– принципу виділення структурних одиниць (розпізнавання за шаблоном, виділення лексичних елементів).

На рисунку 1.1 представлена класифікація систем розпізнавання мови.

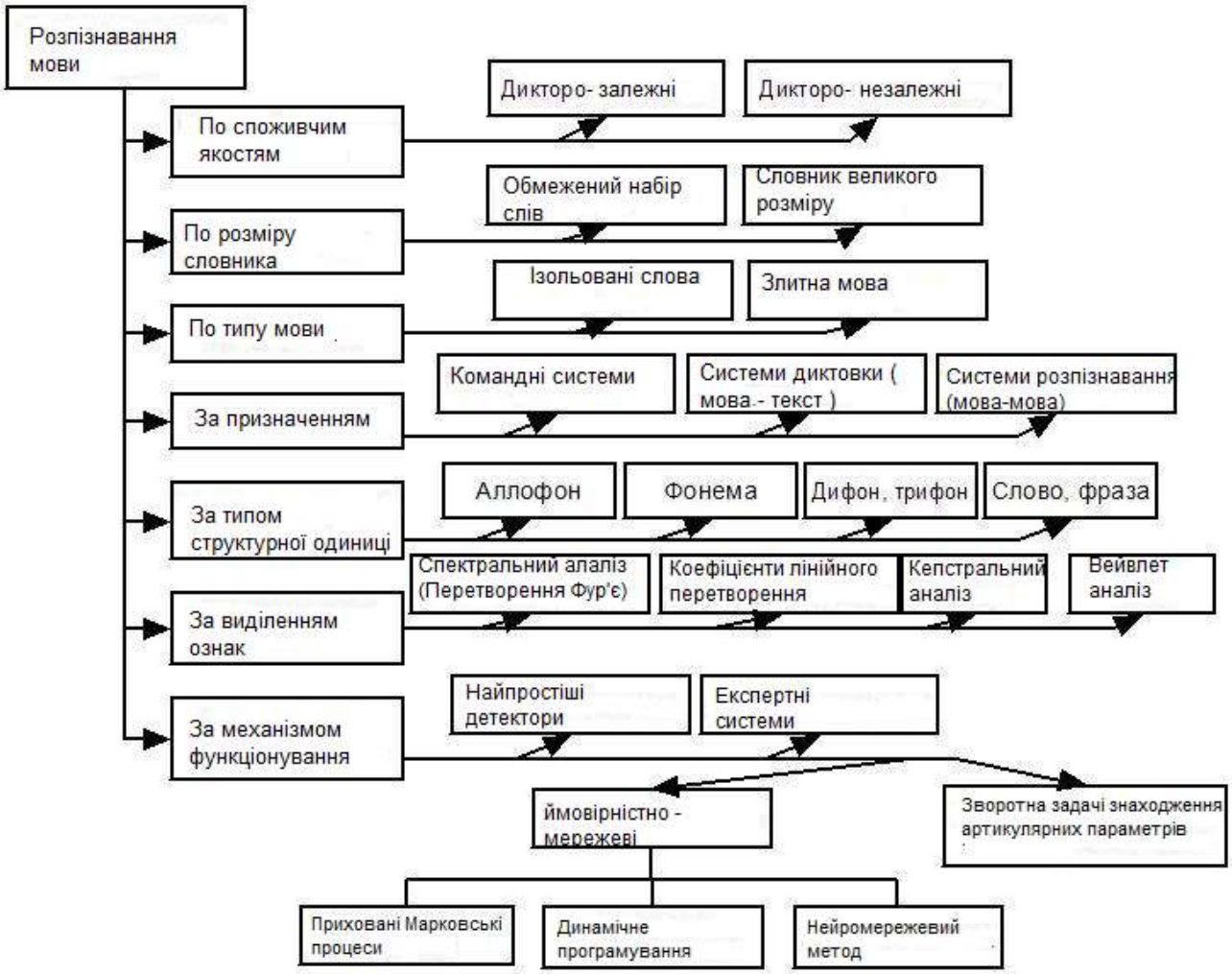

Рисунок 1.1 – Класифікація систем розпізнавання мови

За типом структурної одиниці. При аналізі мови, в якості базової одиниці можуть бути обрані окремі слова або частини вимовлених слів, такі як фонеми, дифон або трифон, алофон. Залежно від того, яка структурна частина обрана, змінюється структура, універсальність і складність словника.

За виділенням ознак. Сама послідовність відліків тиску звукової хвилі – надмірно надлишкова для систем розпізнавання звуків і містить багато зайвої інформації, яка при розпізнаванні не потрібна, або навіть шкідлива. Таким чином, для подання мовного сигналу з нього потрібно виділити будь-які параметри, адекватно представляють цей сигнал для розпізнавання.

За механізмом функціонування. У сучасних системах широко використовуються різні підходи до механізму функціонування систем, що розпізнають. Ймовірносно-мережевий підхід полягає в тому, що мовний сигнал розбивається на певні частини (кадри, або по фонетичному ознакою), після чого відбувається імовірнісна оцінка того, до якого саме елементу розпізнається словника має відношення дана частина і (або) весь вхідний сигнал. Підхід, заснований на вирішенні оберненої задачі синтезу звуку, полягає в тому, що по вхідному сигналу визначається характер руху артикуляторов мовного тракту і, за спеціальним словником відбувається визначення вимовлених фонем.

Цінність голосових технологій для біометрії була неодноразово доведена. Однак тільки висока якість реалізації автоматичних систем розпізнавання диктора здатна реально впровадити такі технології в практику. Подібні системи вже існують. Вони знаходять застосування в системах безпеки, в банківських технологіях, електронної комерції, правоохоронній практиці.

Використання систем розпізнавання диктора є найбільш природним і економічним способом вирішення проблем несанкціонованого доступу до комп'ютера або системам передачі інформації, а також проблем багаторівневого контролю доступу до мережевих або інформаційних ресурсів.

Системи розпізнавання диктора можуть вирішувати два завдання: визначати особистість із заданого, обмеженого списку людей (ідентифікація особистості) або підтверджувати особистість мовця (верифікація особистості). Ідентифікація та

верифікація особистості по голосу є напрямками розвитку технології обробки мовлення.

В даний час розпізнавання мови зводиться до вирішення трьох типів завдань:

– розпізнавання окремо вимовлених слів (використовується для мовного управління обчислювальною машиною);

– розпізнавання злитого мовлення (має на меті перетворення в текст природної мови людини);

– ідентифікація за зразком мови (використовується для цілей забезпечення безпеки).

У процесі реєстрації користувача запам'ятовуються особливості його голосу і формується так звана мовна модель. При тестуванні виконується порівняння запропонованого зразка мови з запомненной мовної моделлю користувача, а також з моделлю "самозванця", складеної на базі голосів безлічі інших людей. Якщо результат порівняння виявиться позитивним для першого випадку і негативним для другого, вважається, що тестування пройшло успішно.

На рисунку 1.2 представлені напрямки розвитку технологій розпізнавання диктора.

Рисунок 1.2 – Напрямки розвитку технологій розпізнавання диктора

1.2 Огляд підходів і технологій ідентифікації голосу

Наступні дати можна назвати основними віхами в розвитку комп'ютерного розпізнавання мовлення.

1962 р. – перший комерційний пристрій мовного виводу: модель 7772 від IBM.

1984 р. – перша система розпізнавання мови на базі ЕОМ. На розпізнавання слова йшли хвилини. Система розрізняла приблизно 5000 слів.

1986 р. – дослідний зразок системи мовного вводу Tangora 4. Завдяки спеціальному мікропроцессору вперше стала можлива обробка мови на робочому місці в реальному часі. В системі вже з'явилася функція контролю контексту.

IBM Tangora, названа на честь Альберта Тангора, стала найшвидшою друкаркою в світі, і могла пристосуватися до голосу мовця. Це як і раніше вимагало повільної, чіткої мови і відсутності фонового шуму, але використання прихованих моделей Маркова дозволило підвищити гнучкість завдяки кластеризації даних і передбачення майбутніх фонем на основі попередніх патернів.

Хоча для кожного користувача знадобилося 20 хвилин навчання (у формі записаної мови), Тангора могла розпізнати до 20 000 англійських слів і кілька повних пропозицій.

1990 р. – Dragon System представила першу американську версію програми мовного введення Dragon Dictate System.

1992 р. – технологія Tangora в моделі клієнт-сервер. Використовується RISC-система IBM RS / 6000. Мовне введення з ПК під OS / 2.

1993 р. – з'явилася перша система мовного вводу для ПК – Personal Dictation від IBM. Одночасно виходить Philips Dictation System – перша система безперервного розпізнавання мови.

1995 р. – IBM представила на CeBIT систему диктовки VoiceType зі спеціалізованими словниками для медиків і адвокатів.

1997 р. – з'явилася система клієнт-сервер Speech Magic від Philips. Lernout & Hauspie представила першу англомовну систему розпізнавання мовлення [5].

2001 р. – Microsoft випускає комплект офісних додатків Office XP з підтримкою мовного вводу й управління.

2002 р. – Google запускає, правда в тестовому режимі, Voice Search, призначеного для голосового пошуку в мережі інтернет. Але дану розробку довелося відразу згорнути. Справа в тому що, що б виконувати даний пошук, потрібно дзвонити на спеціальний номер, що було дуже незручно. Але Google не опустив рук, і продовжував розробки в цьому напрямі.

2005 р. – виходить перша операційна система з функцією розпізнавання мовлення. Першовідкривачем була Mac OS X Tiger. Однак слід згадати, що подібні напрацювання були і у Windows 95, але там була швидше тестова версія, ніж повноцінний продукт. VoiceOver була здатна не тільки на розпізнання мовлення, на додаток до цього вона була її синтезатором. Ця програма могла прочитати вміст текстових документів, поштових і веб-сторінок. Великим плюсом було те, що вона була спікеронезавісімой, і навіть працювала з кількома користувачами одночасно.

2006 р. – Microsoft випускає операційну систему з повноцінною підтримкою функції розпізнавання мовлення Windows Vista.

2009 р. – виходить додаток Voice Search від Google для iPhone. Робота цього додатка спирається на захмарні обчислення своїх суперкомп'ютерів. Ці обчислення дозволили провести великомасштабний аналіз даних пошуку збігів між величезним числом голосових запитів користувачів і їх словами. Ця процедура сприяла швидкому зростанню і вдосконаленню системи. Voice Search поступово закріплює за собою славу самого популярного додатка від Google для мобільних пристроїв. З'являється версія для Android.

Завдяки великому обсягу даних для навчання, додаток Voice Search продемонструвало чудові поліпшення точності в порівнянні з попередніми технологіями розпізнавання мовлення. Google ввів елементи персоналізації в свої результати пошуку за допомогою голосу, і використовував ці дані для розробки свого алгоритму Hummingbird, отримуючи набагато більш тонке розуміння використовуваної мови. Ці нитки були пов'язані в Google Assistant, який зараз є майже на 50% всіх смартфонів.

2011 р. – Google врахував помилки минулих років, результатом чого стала функція розпізнавання голосу в браузері Chrome. Були усунені непотрібні дзвінки та інші незручності. На сьогоднішній день в базі налічується близько 230 мільярдів слів на багатьох мовах світу [3].

Розвиток галузі знань, пов'язаної з аналізом і розпізнаванням мовного сигналу, почалося з вирішення завдань передачі мови по вузькосмуговим каналам у зв'язку з пропускною здатністю меншою, ніж у звичайній телефонній лінії. Вирішення цього завдання привело до створення вокодеров – пристроїв, які виконують скорочення частотної смуги мовних сигналів для ліній телекомунікації. Першим успіхом в цій галузі вважається смуговий вокодер американського інженера-зв'язківця Х. Дадлі. Він представляв собою параметричний вокодер, фільтрував спектр мови з інтервалом в 20-30 мс на кілька смуг, в кожній з яких вимірювалася енергія. Вокодер спочатку здійснював спектрально-часовий аналіз мовного сигналу, виділяючи його акустичні параметри, а потім міг відновити (ресинтезувати) вихідний мовний сигнал на підставі виділених параметрів. На відміну від попередніх сінтезаторів, вокодер Дадлі був заснований не на імітації артикуляції, а на відтворенні акустичних параметрів мовного сигналу.

Серйозні роботи з розпізнавання мовлення почались в основному після Другої світової війни. Перший пристрій для розпізнавання мовлення з'явився у 1952 році, він міг розпізнавати вимовлені людиною цифри. У AT & T Bell Labs була створена система розпізнавання окремих цифр за допомогою простого узгодження акустичних характеристик з шаблонами. Вона представляла собою досить примітивну систему, яка могла розпізнавати цифри, передані голосом по телефону.

Для подальшого розвитку автоматичного розпізнавання мовлення (АРР), велике значення мали метод динамічної спектрографії (типу "Видима мова") і широке використання відповідної апаратури в фонетичних дослідах. До кінця 50-х років на матеріалі самих різних мов був накопичений великий досліднецький матеріал, який свідчив про складну природу відповідності між прівичними для лінгвістів уявленнями мовних відрізків у вигляді послідовності фонем або аллофонів і фізичною реальністю звукової мові.

На початку 60-х років компанія IBM розробила і продемонструвала "Shoebox" – попередника сучасних систем розпізнавання мовлення. Це новаторський пристрій розпізнавав і реагував на 16 вимовлених слів, включаючи цифри від 0 до 9. Воно було показано по телебаченню і в павільйоні IBM на світовій ярмарку 1962 року в Сіетлі.

Система розпізнавання на основі вірогіднісного підходу була створена Фраєм і Денесом в лондонському University College. У цій системі вперше використовувалися ймовірності переходів між фонемамі. Починаючи з 1971 р Агентство перспективних дослідницьких програм (DARPA) міністерства оборони США фінансувало чотири конкуруючі п'ятирічні проекти по розробці високоефективних систем розпізнавання мови. Переможцем цієї програми і єдиною системою, що відповідає вимогам по розпізнаванню словника з 1000 слів з точністю 90%, стала система HARPY, розроблена в університеті CMU. Остаточна версія цієї системи була створена на основі системи Dragon, розробленою аспірантом того ж університету Дж.Бейкером. У цій системі для імовірнісного моделювання слів мови вперше були використані приховані марковскі моделі. Прихована марківська модель є на сьогоднішній день найбільш широко застосовуваним і ефективним підходом до проблеми побудови акустичної моделі.

Майже одночасно з системою Dragon в компанії IBM була розроблена ще одна система на основі прихованих марківських моделей. Починаючи з цих двох розробок, імовірнісні методи у цілому і приховані марковскі моделі зокрема стали домінувати в дослідженнях і розробках по розпізнаванню мовлення. Використання даного підходу, зважаючи на свою ефективність, стало в теперешній час майже промисловим стандартом.

У 1980-х завдяки новим підходам і технологіям словниковий запас подібних систем виріс з декількох сотень до декількох тисяч слів і мав потенціал розпізнавання необмеженої кількості слів. Однією з причин був новий статистичний метод, більше відомий як прихована марківських модель.

Використовуючи шаблони для слів і звукові патерни, вона розглядала ймовірність того, що невідомі звуки могли бути словами. Ця база використовувалася

іншими системами ще протягом двадцяти років (Automatic Speech RecognitionA Brief History of the Technology Development).

З розширеним словниковим запасом розпізнавання мови початок торувати собі доріжку в комерційні додатки для бізнесу та спеціалізованих галузей, таких як медицина.

У дев'яностих комп'ютери нарешті отримали швидкі процесори, і програми з розпізнавання мови стали життєздатними.

В 1990 році з'явилася перша загальнодоступна програма Dragon Dictate с приголомшливою ціною 9000 доларів. Через сім років вийшла поліпшена версія – Dragon NaturallySpeaking. Додаток распознавало нормальну мову, тому ви могли говорити в звичайному темпі близько 100 слів за хвилину. Але все одно, ви повинні були тренувати програму протягом 45 хвилин перед використанням, і вона мала все ще високу ціну в 695 доларів.

Поява першого голосового порталу VAL від BellSouth було в 1996 році. Це була перша інтерактивна система розпізнавання мови, яка давала інформацію, грунтуючись на тому, що ви сказали в трубку телефону.

До 2001 року розпізнавання мовлення піднялося до 80-відсоткової точності, і прогрес технології зупинився. Системи розпізнавали працювали відмінно, коли мовна всесвіт була обмеженою, але вони до сих пір «здогадувалися» за допомогою статистичних моделей серед схожих слів, мовна всесвіт росла разом із зростанням Інтернету.

Розпізнавання голосу і голосові команди був вбудовані в Windows Vista і Mac Os. Більшість користувачів навіть не здогадувалися, що така функціональність існує. Windows Speech Recognition і голосові команди OS X були цікавими, але недостатньо точними і зручними, як клавіатура і миша.

Технологія розпізнавання мовлення отримала друге дихання після одного важливого події: появи додатки Google Voice Search для iPhone. Вплив цього додатка було значним з двох причин. По-перше, телефони та інші мобільні девайси – це ідеальні об'єкти для розпізнавання мовлення, і бажання замінити крихітні екранні клавіатури альтернативними методами введення було дуже велике. По-друге, у

Google була можливість розвантажити цей процес, використовуючи свої хмарні дата-центри, направивши всю їх міць для великомасштабного аналізу даних для пошуку збігів між словами користувачів і величезного числа зразків голосових запитів, які вони отримували.

Якщо коротко, вузьким місцем розпізнавання мовлення завжди було доступність даних і можливість ефективної їх обробки. Додаток ж додало до аналізу дані мільярдів пошукових запитів, щоб краще передбачати, що ж ви сказали.

У 2010 році Google додав персональне розпізнавання в голосовий пошук телефонів під управлінням Android. Програмне забезпечення могло записувати голосові запити користувачів для побудови більш точної голосової моделі. Також компанія додала розпізнавання голосу в свій браузер Chrome в середині 2011 року. Система Google тепер дозволяє розпізнати 230 мільярдів слів.

Потім з'явилася Siri. Так само, як і система Goggle Voice Search, вона покладається на хмарні обчислення. Вона використовує ті дані, які їй відомі про тебе, щоб згенерувати випливає з контексту відповідь і відповідає на твій запит, як якась особистість[4].

Зростання ринку розпізнавання голосу (верифікації диктора) по голосу пов'язаний безпосередньо з цілою низкою чинників. Головним із чинників є збільшення ваги на голосовий біометрії, в зв'язку з пошкодженням (і ускладненням) порушення безпеки, тому що Безпека продовжує залишатися головним з основних вимог для різних (в тому числі і державних). Голосова біометрія має вирішальне значення в ідентифікації, тому що голос має унікальні характеристики для будь-якої людини. Саме тому даний вид біометричної ідентифікації має великий попит.

Деякі з основних чинників світового ринку розпізнавання мовлення та ідентифікації диктора:

–підвищення попиту на послуги голосової біометрії;

–більш широке використання диктора для судово-медичних цілей, криміналістики, електронної комерції, банківських технологій і багатьох сферах життєдіяльності;

–попит на розпізнавання мовлення у військових цілях;

–зацікавленість технологіями call-центрами;

–високий попит для розпізнавання голосу в сфері охорони здоров'я.

Перевагою голосової біометрії є простота реалізованої системи, як правило, складається з голосового приймача, диктофона, голосового модулятора, біометричного програмного забезпечення та бази даних. Фактично система розпізнавання голосу може бути реалізована на базі ПК.

Крім того, на відміну від інших біометричних технологій, голосова біометрія дозволяє здійснювати верифікацію на великій відстані.

Головну рушійну силу розвитку даної технології на світовому ринку формує її широке застосування в судовій сфері медицини. Відбувається складний процес аналізу і визначення, чи відповідає набір характеристик голосу людини, підозрюваного в скоєнні злочину, голосу з бази судово-медичних зразків. Ця технологія дозволяє правоохоронним органам виявляти злочинців по одній з найбільш унікальних характеристик людини, тим самим пропонуючи високий рівень точності. Судово-медичні експерти проводять аналіз голосу підозрюваного зразкам того часу, поки не буде знайдений злочинець. Останнім часом ця технологія використовується, щоб допомогти вирішити деякі кримінальні справи.

Для запобігання проникнення зловмисників використовують сучасні технології розпізнавання диктора за голосом для запобігання проникнення зловмисників. Дані біометричні системи дозволяють виявляти наявність несанкціонованих проникнень в захищену зону. Система база даних голосів тих осіб, які мають допуск до секретних даних або території.

Ці люди проходять процес ідентифікації (рідше верифікації) системою розпізнавання особистості, найвища можливість отримання доступу тим людям, чиїх голосів немає в базі даних системи. Таким чином, дані тези зумовлюють високий попит на ідентифікацію диктора за голосом для військових цілей.

Зокрема, система верифікації особистості по голосу, як очікується, стане одним з головних критерієм у встановленні особи в медичних установах. Багато компаній охорони здоров'я в США, звертаючись до стандартів Закон про переносимості та підзвітності медичного страхування (HIPA), також використовує біометричні

технології, такі як розпізнавання голосу, розпізнавання відбитків пальців для більш безпечної та ефективної реєстрації пацієнта, накопичення інформації пацієнта, захисту медичних записів пацієнта. Також установи клінічних випробувань впроваджують розпізнавання голосу для нинішніх осіб, набраних для клінічних випробувань. Таким чином, голосова біометрія є одним з основних методів ідентифікації в системі охорони здоров'я в Азіатсько-Тихоокеанському регіоні.

Аналітики TechNavio прогнозують глобальне зростання систем розпізнавання голосу: в середньому ринок виросте на 22,15% протягом 2014-2019 років [16].

1.3 Постановка задачі розробки методу аутентифікації користувача через голосове повідомлення

Задача яка вирішується в рамках атестаційної роботи – розробка методу аутентифікації користувача через голосове повідомлення. Цільова аудиторія користувачів такого програмного засобу – системи розмежування доступу, при проведенні фінансових транзакцій, при аутенифікації в особистих пристроях і т.д.

Для спеціалістів в області голосового розпізнавання важливо забезпечити максимальний відсоток правильної веріфікації при мінімальних зусиллях з боку користувача програмою. Для цього використовують модуль шумоочистки і відділення корисного сигналу, акустичну та язикову моделі. Для розробки методу буде розглянутий алгоритм динамічного викривлення. Необхідно також знайти спосіб покращення цього алгоритму.

Розроблюваний метод ідентифікації голосу повинен виконувати наступні задачі:

– оцінка якості мовного сигналу. На цьому етапі визначається рівень перешкод і спотворень;

– оцінка інформації про частини мови, форми слова і статистичні зв'язки між словами;

– ідентифікація голосу;

– провести аналіз результатів тестування програмного засобу та встановити, наскільки вихідні дані програми задовільняють критеріям виконання поставленої в рамках роботи задачі.

Вхідною інформацією є голосове повідомлення довжиною максимум п’ять секунд, достатньо одного слова. На виході буде виведено повідомлення про результат ідентифікації – іденфіковано/не ідентифіковано.

Результатом атестаційної работи буде математичний метод, який можно використовувати для ідентифікації диктора за ключовим словом.

Для аналізу повноти реалізації функцій програмного засобу повинно бути проведене тестування ідентифікації голосу. Будуть відібрані голосові сигнали декількох чоловік. Записи мови проводилися в моно режимі за допомогою вбудованого в комп'ютер мікрофона, що має частоту дискретизації 16 кГц. Тривалість мовного сигналу становила 3 секунди. Критерій виконная здачі – критерії false acceptance rate (FAR) і false rejection rate (FRR):

$$FAR = \frac{N_{FAR+}}{N_{FAR\ all}}, \tag{1.1}$$

де FAR – частка помилкових підтвердження користувача; N_{FAR+} – кількість успішних автентифікацій зловмисника; $N_{FAR\ all}$ – загальна кількість спроб пройти автентифікацію зловмисником.

$$FRR = \frac{N_{FRR-}}{N_{FRR\ all}}, \tag{1.2}$$

де FRR – частка помилкових відмов користувачеві; N_{FRR-} – кількість відмов у автентифікації користувача; $N_{FAR\ all}$ – загальна кількість спроб пройти автентифікацію користувачем.

Одним з критеріїв роботи системи може бути підхід, що полягає в наступному: система тим краще, чим менше значення FRR при однакових значеннях FAR.

На даний момент існує безліч компаній (Agnitio, Nuance, Voice Security Systems), що розробляють системи голосової біометрії. У більшості розроблених систем ймовірність помилки ідентифікації становить 1 - 3%, але дані розробки мають ряд недоліків.

Слід звернути увагу на наступні дослідження:

- Agnitio (FAR: 2.1%, FRR: 1.5%);
- Nuance (FAR: 2.6%, FRR: 1.8%);
- Voice Security Systems (FAR: 1.1%, FRR: 2.2%);
- VoiceTrust (FAR: 2.5%, FRR: 2.8%);

Необхідно розуміти, що метод розробляється для повсякденного користуванняя, тобто юдина, яка проходить тест, є користувачем мобільного пристрою або ПК, отже, така ситуація накладає ряд істотних обмежень.

По-перше, користувач сильно чутливий до швидкості отримання доступу, тобто немає можливості для тривалих тестів і складних процедур. Так само FRR буде значно важливіше, ніж в інших умовах.

По-друге, найчастіше, користувачів мобільного пристрою вельми обмежена кількість. Отже, значимість FAR трохи слабше, ніж при інших умовах.

2 ОПИС ТА АНАЛІЗ ПІДХОДІВ ДО РОЗРОБКИ ПРОГРАМНИХ ЗАСОБІВ ДЛЯ РОЗПІЗНАВАННЯ ГОЛОСОВИХ КОМАНД

2.1 Опис алгоритму розпізнавання голосових команд

На рисунку 2.1 наведний загальний алгоритм розпізнавання голосових команд.

Рисунок 2.1 – Загальний алгоритм розпізнавання голосових команд

Цей загальний алгоритм застосовується для всіх засобів розпізнавання мовлення. Попередня корекція та фільтрація і подавлення шумів не є обов’зковим етапом для розпізнавання мовлення.

Розглянемо більш докладно систему верифікації диктора. Верифікація особистості по голосу – це визначення, чи є хто говорить тим, ким він є. Користувач, раніше зареєстрований в системі, вимовляє свій ідентифікатор, який представляє собою реєстраційний номер, парольне слово або фразу. При текстозавісімом розпізнаванні парольне слово відомо системі, і вона «просить» користувача вимовити його. Парольне воно виглядає неправильним, і людина вимовляє його в мікрофон. При текстонезавісімом розпізнаванні промовлене користувачем парольне слово не збігається з еталонним, тобто в якості пароля користувач може вимовляти довільне слово або фразу. Система верифікації приймає мовний сигнал, обробляє його і

вирішує, прийняти або відхилити пропонований користувачем ідентифікатор. Система може повідомити користувачеві про недостатню ступеня збігу його голосу з наявним еталоном і попросити вимовити додаткову інформацію, щоб прийняти остаточне рішення.

Користувач вимовляє в мікрофон пропонований йому системою номер для того, щоб система перевірила, чи відповідає його голос еталону, що зберігається в базі даних системи. Як правило, існує компроміс між точністю розпізнавання голосу і розміром мовного зразка, тобто чим довше мовної зразок, тим вище точність розпізнавання. Крім голосу в мікрофон можуть потрапляти відлуння і сторонні шуми.

Схема взаємодії людини з системою верифікації особистості по голосу зображена на рис. 2.2.

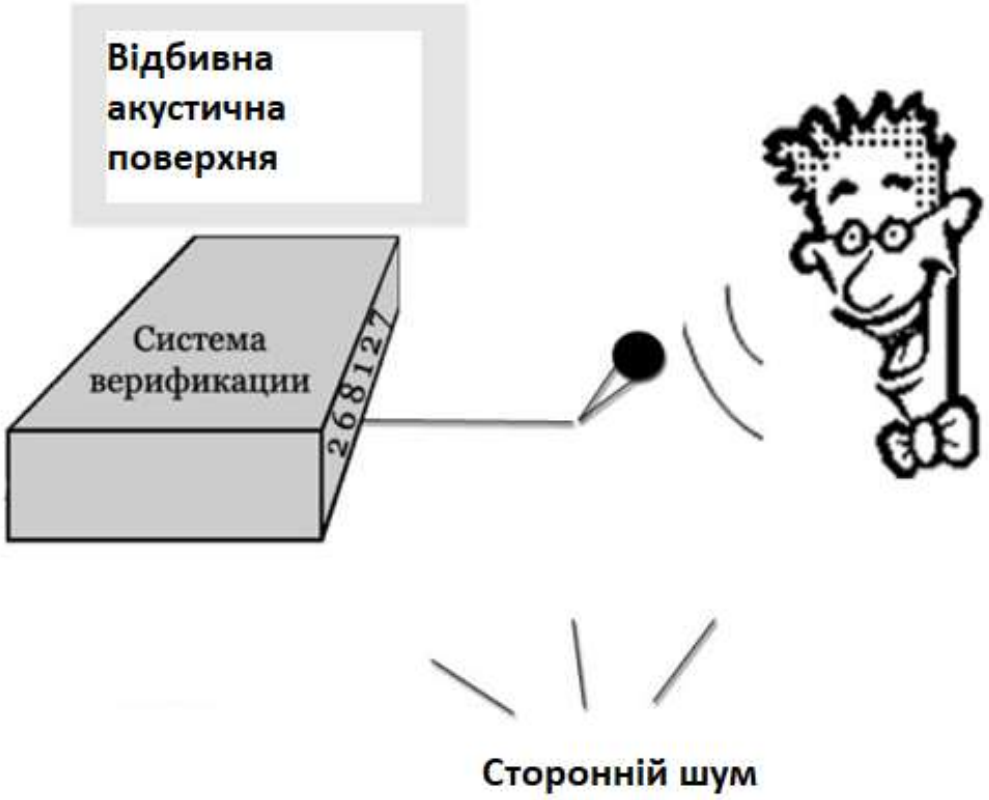

Рисунок 2.2 – Схема взаємодії людини з системою верифікації особистості по голосу

Існує ряд факторів, які можуть сприяти виникненню помилок верифікації та ідентифікації, наприклад:

– неправильне проголошення або прочитання парольного слова або фрази;

– емоційний стан диктора (стрес, проголошення парольної фрази під примусом та ін.);

– складна акустична обстановка (шум, перешкоди, радіохвилі тощо.);

– різні канали зв'язку (використання різних мікрофонів під час реєстрації диктора і верифікації);

– простудні захворювання;

– природні зміни голосу.

Деякі з них можуть бути усунені, наприклад, шляхом використання більш якісних мікрофонів. На рисунку 2.3 зображений загальний алгоритм веріфікації диктора.

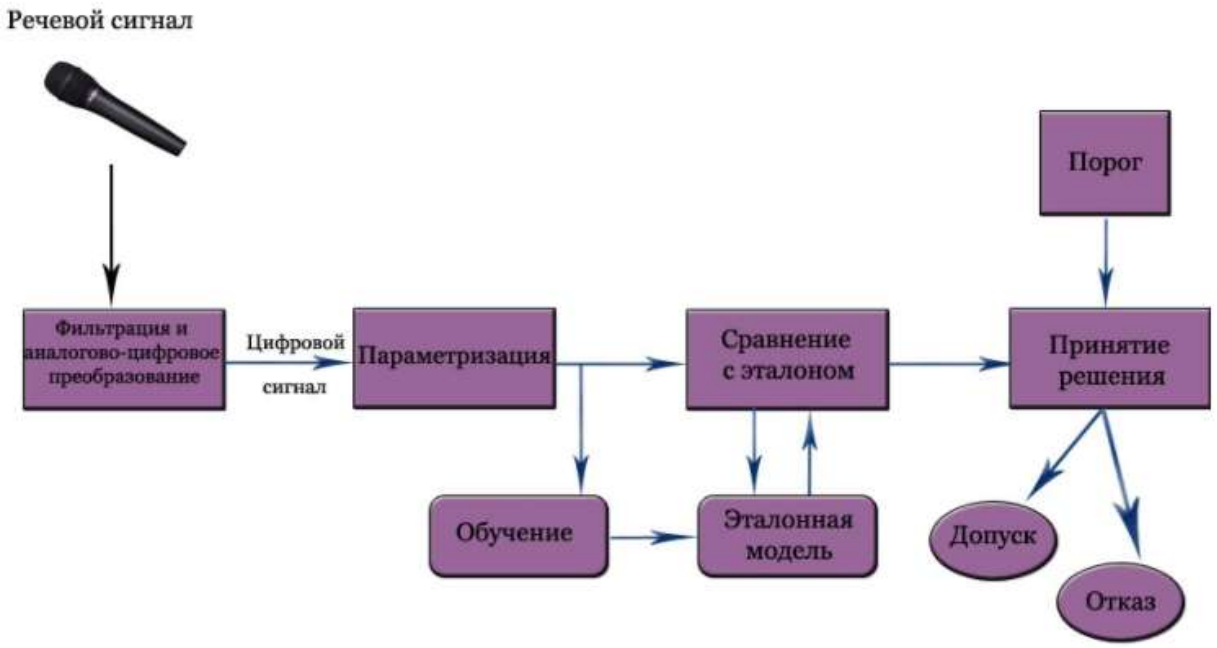

Рисунок 2.3 – Загальний алгоритм веріфікації диктора

Під час реєстрації новий користувач вводить свій ідентифікатор, а потім вимовляє кілька разів ключове слово або фразу, таким чином створюються еталони. Число повторів ключової фрази може варіюватися для кожного користувача, а може бути постійним для всіх.

Для того щоб комп'ютер міг обробити мовний сигнал, звукова хвиля перетворюється в аналоговий, а потім в цифровий сигнал.

На етапі виділення ознак голосу мовний сигнал розбивається на окремі звукові кадри, які згодом перетворюються в цифрову модель. Ці моделі називають «голосовими відбитками». Знову отриманий «голосовий відбиток» порівнюється з раніше встановленим еталоном. Для розпізнавання особистості говорить найважливішими є найбільш яскраві відмінні ознаки голосу, які дозволили б системі з високою точністю розпізнавати голос кожного конкретного користувача.

Нарешті, система приймає рішення допустити або відмовити користувачеві у допуску в залежності від збігу або розбіжності його голосу з встановленим еталоном. Якщо система невірно зіставили пред'явлений їй голос з еталоном, то виникає помилка «помилковий допуск» (FA). Якщо ж система не впізнала біометричний ознака, який відповідає наявному в ній ідеалу, то говорять про помилку «помилковий відмова» (FR). Помилка помилкового допуску створює пролом в системі безпеки, а помилка помилкового відмови призводить до зменшення зручності користування системою, яка іноді не розпізнає людини з першого разу. Спроба знизити ймовірність виникнення однієї помилки призводить до більш частого виникнення іншої, тому в залежності від вимог до системи вибирається певний компроміс, тобто встановлюється поріг прийняття рішення.

Процес верифікації особистості по голосу складається з 5 етапів: прийом мовного сигналу, параметризація, або виділення характерних ознак голосу, порівняння отриманого зразка голосу з раніше встановленим еталоном, прийняття рішення «допуск / відмова», навчання, або оновлення еталонної моделі.

2.2 Опис алгоритму оцінки якості голосової команди

Обробка мовних сигналів – це область науки, в якій здійснюються фільтрація і придушення шумів, посилення, поділ інформаційних потоків, вилучення інформації, кодування, стиснення і відновлення мовних сигналів. Вона отримала широке поширення у всіх напрямках мовних технологій.

Чіткість голосу – відносна кількість правильно прийнятих елементів мови (звуків, складів, слів, фраз), виражене у відсотках від загального числа переданих елементів.

Якість мови – параметр, що характеризує суб'єктивну оцінку звучання мови в випробуваної системі передачі мови.

Нормальний темп мови – проголошення промови зі швидкістю, при якій середня тривалість контрольної фрази дорівнює 2,4 с.

Прискорений темп мови – проголошення промови зі швидкістю, при якій середня тривалість контрольної фрази дорівнює 1,5 – 6 с.

Впізнаваність голосу мовця – можливість слухачів ототожнювати звучання голосу, з конкретною особою, відомим слухачеві раніше.

Смислова розбірливість – показник ступеня правильного відтворення інформаційного змісту промови.

Інтегральне якість – показник, що характеризує загальне враження слухача від прийнятої мови [6].

Напрямок обробки мовних команд в системах голосового управління включає наступні завдання:

– реєстрацію;
– попередню корекцію;
– фільтрацію і придушення шуму;
– сегментацію на фрейми;
– сегментацію «сигнал / пауза»;
– сегментацію «тон / НЕ тон»;
– визначення інформативних параметрів.

Фільтрація і придушення шуму – це етап обробки мовних команд, який дозволяє підвищити розбірливість, зменшити частку шумів, викликаних як акустичними, так і технологічними причинами.

На рисунку 2.4 описані мовні технології.

Рисунок 2.4 – Мовні технології

Шум – безладні коливання різної фізичної природи, що відрізняються складністю тимчасової і спектральної структур. Стосовно до мовним сигналам шум – цей сукупність апериодических звуків різної інтенсивності і частоти, які змінюють інформативні параметри сигналу.

Шуми із взаємодії з корисним мовним сигналом діляться на адитивні та мультиплікативні. Адитивні шуми складаються з корисним сигналом і вносять незначну похибку. Мультиплікативні шуми перемножуються з корисним сигналом і вносять найбільшу похибка – можуть змінювати інформативні параметри мовних команд. У найзагальнішому вигляді комбінація сигналу і шуму виглядає наступним чином:

$$S(t) = \left(k_c(t) + k_{\text{ш}}(t)\right) * e(t) + n(t), \qquad (2.1)$$

де $e(t)$ – корисний мовний сигнал; $k_c(t)$ – коефіцієнт, що характеризує корисний мовний сигнал; $k_{\text{ш}}(t)$ – коефіцієнт, що характеризує мультиплікативний шум; $n(t)$ – адитивний шум

Акустичні мовні сигнали часто спотворені адитивними перешкодами, значно знижують ефективність систем верифікації диктора. У загальному випадку дані адитивні перешкоди можуть бути розділені на дві великі групи: стаціонарні, присутні на всьому протязі сигналу (наприклад, широко відомий білий і рожевий шум), і нестаціонарні короткочасні, присутні на окремих ділянках сигналу.

Рисунок 2.5 – Класифікація шумів у мовних сигналах

При наявності перешкод другої групи вхідні сигнали рідко бувають повністю спотворені. Незначно спотворені ділянки сигналу чергуються з ділянками, сильно спотвореними імпульсними перешкодами різних типів: Кліпування, короткочасними

електричними наведеннями, перевантаженнями і т.п. Саме ці нестаціонарні перешкоди і спотворення мають найбільший негативний вплив. Відповідно використовуючи детектори, здатні на етапі попередньої обробки з високою ймовірністю виявляти подібного роду перешкоди і спотворення (з метою їх подальшого придушення або виключення з аналізу), можна істотно поліпшити якість систем обробки мови [7].

Відношення інтенсивності сигналу I_c і шуму $I_{\text{ш}}$. Це відношення називається «ставлення сигнал / перешкода» і грає важливу роль в завданні фільтрації і шумозаглушення. Відношення сигнал / перешкода виражається в логарифмічних безрозмірних одиницях – децибелах (dB, дБ):

$$N = 10lgI_c/I_{\text{ш}}, \qquad (2.2)$$

де I_c, $I_{\text{ш}}$. – інтенсивності сигналу і шуму.

На рисунку 2.5 зображена класифікація шумів у мовних сигналах.

За походженням шуми в мовних командах можна розділити на фізіологічні і антропогенні. До першого виду шумів відносяться комплекс звуків різної інтенсивності і частоти, що знаходяться в безладному поєднанні з корисними мовними сигналами.

Походження цього виду шумів безпосередньо пов'язано з порушеннями мови (порушення роботи окремих або комплексу органів артикуляції відділу мовного апарату). Наука, що вивчає порушення мови, їх подолання та попередження засобами корекційного навчання називається логопедией. До шумів, пов'язаних з порушенням мови, відносять велику кількість звуків, форма і структура яких безпосередньо пов'язана з родом порушення звуковимови:

– порушення темпу і ритму мовних сигналів (браділалія, тахілалія, спотикання, заїкання);

– порушення голосу (афонії, дисфонии, рінофонії);

– розпад мовних сигналів (афазія).

До антропогенним шумів в грубій інтерпретації, крім фізіологічних, відносяться всі інші види шумів. Назва «антропогенний» походить від зв'язку з людиною, іншими словами, це шуми, що походять від людини і що виникли в результаті його діяльності. Їх також називають промисловими або виробничими шумами. Джерелами антропогенних шумів є транспортні засоби: автомобілі, залізничні потяги і літаки, промислові підприємства, будівельні та ремонтні роботи, побутова і офісна техніка і т.д.

За постійності параметрів все шуми підрозділяються на стаціонарні та нестаціонарні. Стаціонарний шум – шум, який характеризується постійністю середніх параметрів: інтенсивності (потужності), розподілу інтенсивності по спектру (спектральна щільність), автокореляційної функції. Класичною моделлю стаціонарного шуму є білий шум, спектральні складові якого рівномірно розподілені по всьому діапазону задіяних частот.

Основними типовими перешкодами і спотвореннями, розглянутими в цій статті, є клацання, перевантаження, короткі тональні сигнали, кліппування.

У виявленні клацань є певні труднощі, оскільки короткі імпульси, які сприймаються людиною на слух як "клацання", можуть в загальному випадку істотно відрізнятися як в часовому, так і в частотному поданні (на рисунку 2.6 зображені види клацань, 1 – короткий "класичний "високочастотний клацання; 2 – низькочастотне клацання; 3 – клацання з короткими осцилляціями; 4 – довге клацання з шумовим заповненням).

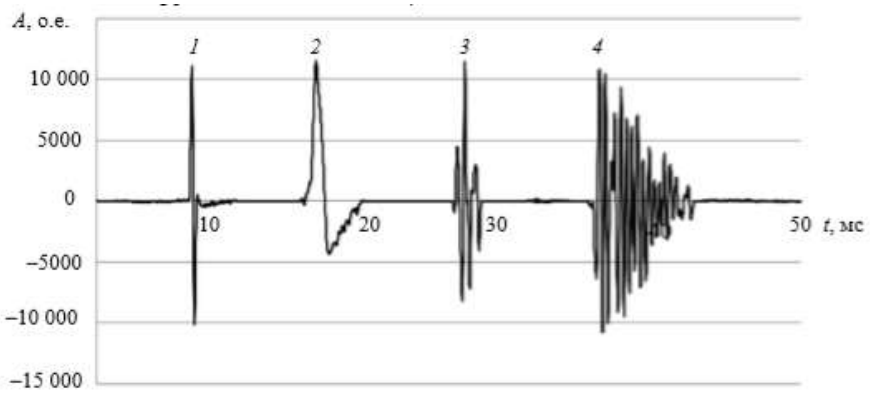

Рисунок 2.6 – Види клацань

Наприклад, короткий високочастотний клацання добре виявляється в такий спосіб. Аналізований сигнал x (i), де i – дискретний тимчасової індекс, спочатку пропускається через високочастотний (ВЧ) фільтр з частотою зрізу близько 2-4 кГц. Потім обчислюється перша різниця d (i) = y (i) – y (i – 1), де y (i) – сигнал на виході фільтра, далі її абсолютна величина порівнюється з граничним значенням.

На жаль, даний спосіб не працює на низькочастотних (НЧ) клацаннях (крива 2), так як, по-перше, основна частина їх енергії зосереджена в низькочастотної області і "зрізається" ВЧ-фільтром, а по-друге, значення d (i) клацань даного виду і мовних сигналів розрізняється несуттєво.

Результати досліджень різних алгоритмів, заснованих на методах лінійного передбачення і авторегресійних моделях показали їх високу обчислювальну складність, тому був розроблений більш простий алгоритм виявлення клацань різних типів.

Розроблений алгоритм включає наступні кроки:

а) Вибирається довжина вікна аналізу (t_0, t_3) таким чином, щоб виконувалася умова $t_3 - t_0 = KL_c$, де L_c – передбачувана тривалість клацання і K – масштабний коефіцієнт, що змінюється в діапазоні від 10 до 100.

б) Вікно розбивається на три частини, довжина центральної частини вибирається сумірною з передбачуваної довжиною клацання, $t_1 - t_0 = t_3 - t_2$.

в) Вихідна величина V_c, порівнювана надалі з граничним значенням, розраховується як:

$$V_c\,(t_{center}) = \frac{2(t_1 - t_0)}{t_2 - t_1} \frac{\sum_{t=t_1}^{t_2} x^2(t)}{\sum_{t=t_0}^{t_1} x^2(t) + \sum_{t=t_2}^{t_3} x^2(t)}\ , \qquad (2.3)$$

де x (t) −аналізований сигнал; center t = 0,5 $(t_3 + t_0)$ – центр інтервалу $[t_3, t_0]$.

Неважко зрозуміти, що V_c в (2.3) є ставлення потужностей сигналу на різних ділянках, нормоване таким чином, що в разі стаціонарного сигналу (наприклад, білого

шуму) $V_c = 1$. Для мовних сигналів отримані значення V_c коливалися від нуля до декількох одиниць.

Величина $V_c > 8$ сигналізує про наявність клацання (строго кажучи, конкретне граничне значення залежить від обраної допустимої ймовірності помилкової тривоги і розмірів вікна аналізу і визначається експериментально).

На рисунку 2.7 зображений розроблений алгоритм виявлення клацань (суцільна крива – ділянка аналізованого сигналу з клацанням, пунктир – вихідна величина алгоритму (помножена на 1000 з метою відображення на одному графіку з сигналом); $t_0 - t_3$ – тимчасові мітки кордонів вікна аналізу).

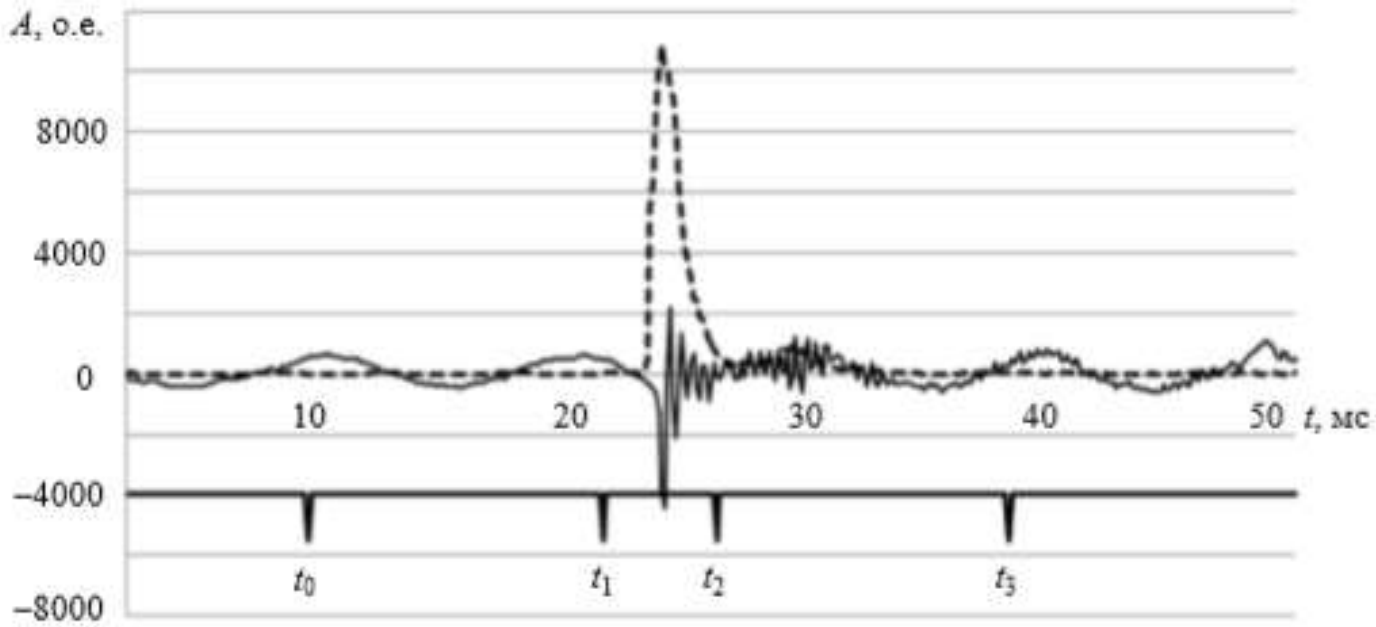

Рисунок 2.7 – Алгоритм виявлення клацань

Очевидно, що довжина інтервалу t2-t1 в ідеальному випадку повинна відповідати тривалості клацання, що підлягає виявленню, що в реальних умовах труднодостижимо. У проведених експериментах встановлено, що якщо це значення знаходиться в межах декількох довжин клацання, то результати детектора також цілком прийнятні. В іншому випадку, при значній апріорної невизначеності в тривалості передбачуваних клацань, доводиться здійснювати перебір.

Шляхом моделювання були отримані наступні тимчасові параметри детектора: інтервал $t_2 - t_1$ 5 мс; $t_1 - t_0$ і $t_3 - t_2$– 60 мс. При таких значеннях отримані хороші результати з детектування типових клацань на реальних мовних сигналах. Слід зауважити, що при виявленні коротких високочастотних клацань буває корисна попередня фільтрація ВЧ-фільтром з частотою зрізу 2-4 кГц.

Перевантаженням називаються короткі (1– 2 відліку) скачки сигналу, імпульси або серії подібних імпульсів великої амплітуди, викликані зміною знака сигналу при так званому "целочисленном переповненні". Причини перевантажень криються в наступному. На практиці найбільш широко використовуваний тип квантування при перекладі аудіосигналів в цифрову форму – 16-бітове квантування. При такому типі квантування кожен відлік сигналу являє собою ціле двухбайтовое число в форматі "signed short int" (стандарт ANSI), тобто амплітуда відліку змінюється від -32 768 до 32 767. У той же час обробка сигналу може виконуватися, наприклад, в форматах "long", "float" або "double". При цьому якщо число, що вийшло після обробки, виходить за межі інтервалу [-32 768 32 767], то при його простому перетвореннi до типу "signed short int" (під час запису, наприклад, на диск в WAV-форматі) відбудеться "перекидання знака ", і число, наприклад 32 768, перетвориться в -32 768 число -32 769 – в 32 767 і т.д.

Загальні вирази для результату можуть бути записані як:

$$\begin{aligned} &\text{if } (x > 32\,767) \text{then } y = (x \bmod 32767) - 32\,768), \\ &\text{if } (x < -32768) \text{ then } y = -(|x| \bmod 32\,768) + 32\,768, \end{aligned} \tag{2.4}$$

де x – число до перетворення, y – результат перетворення, mod – операція обчислення по модулю.

На слух одиночна перевантаження сприймається як високочастотний клацання, а серія подібних клацань – як різкий гучний тріск, істотно погіршує як розбірливість мовного сигналу, так і показники систем обробки мови.

На рисунку 2.8 наведено типовий приклад перевантаження, що виникла при перетворенні величини в форматі double (час перевантаження 6,68 мс, значення x = 56 981) в багатобайтових формат signed short int.

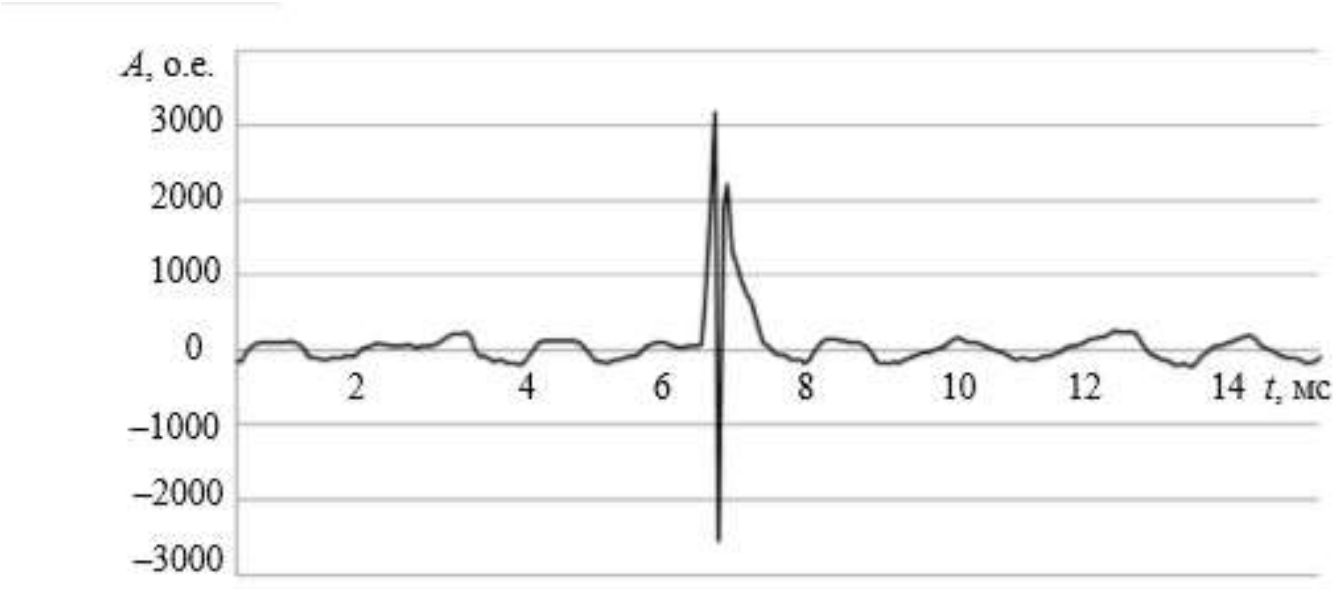

Рисунок 2.8 – Приклад перевантаження

Одиночна перевантаження (на відміну від серії) з успіхом може бути виявлена за допомогою детектора ВЧ-клацань. Однак, використовуючи першу різницю (яка була раніше описана як неефективна при виявленні НЧ-клацань другого типу), можливо створити алгоритм, який виявляє як поодинокі, так і множинні перевантаження. Справа в тому, що "перекидання" знака викликає сильні різкі скачки амплітуди за один відлік, часто сумірні з динамічним діапазоном сигналу. В даному випадку коефіцієнт обчислюється таким чином:

$$d(i) = \frac{|x(i)-x(i-1)|}{A_{max}-A_{min}}, \qquad (2.5)$$

де A_{max} і A_{min} – максимальне і мінімальне значення амплітуди сигналу, обчислені по всій вибірці. Теоретично $0 \leq d(i) \leq 1$, однак на чистої мови, без перевантажень, величина $d(i)$, як правило, значно менше одиниці

Алгоритм детектування перевантажень:

а) Вибирається величина порога T_d, наприклад, 0,7.

б) По всій вибірці сигналу обчислюються його максимальне max A і мінімальне min A значення.

в) Для кожного відліку сигналу x(i) , $i = 1, N - 1$ (N – повна довжина сигналу) за формулою (2.4) обчислюється коефіцієнт d(i). Проводиться порівняння d(i) з обраним раніше порогом, і в разі $d(i) > T_d$ приймається рішення про наявність перевантаження.

Короткі тони – це широко відомі сигнали телефонного виклику, що представляють собою зазвичай одну або дві гармоніки довжиною близько однієї секунди. Відмінною особливістю таких сигналів є високий рівень і стабільність частоти складових гармонік. Відповідно в переважній більшості алгоритмів виявлення тонів використовується аналіз спектрів потужності (або модулів спектрів потужності) сигналів.

Відзначимо, що тональні сигнали без домішки стороннього шуму або в сумі з шумом малої потужності можуть бути також з успіхом виявлені детектором кліппірованних сигналів, що базується на аналізі гістограми.

Алгоритм детектування:

а) Вибирається величина M – довжина сегмента сигналу для обчислення спектру потужності.

б) Для кожного сегмента сигналу довжиною M обчислюється модуль миттєвого спектра потужності S(m), де $m = 0, M/2$ – дискретна частота.

в) Для всіх $m = 0, M/2$ = знаходиться спектральний максимум S_{max}.

г) Обчислюється пороговий рівень $T_s = T_{s0} S_{max}$.

д) Для всіх $m = 0, M/2$ підраховується цільова величина K_s – кількість спектральних відліків, що перевищують рівень T_s, тобто: $K_s = \sum_{m=0}^{M/2} k_s$, де

$$k_s = \begin{cases} 1 \; if \; S(m) \geq T_s \\ 0 \; if \; S(m) < T_s \end{cases} \tag{2.6}$$

е) Проводиться порівняння: якщо $3 \leq K_s$, то приймається рішення про наявність тональної складової в досліджуваному фрагменті сигналу. В алгоритмі оцінки сталості амплітуди спектральних максимумів використовується той факт, що на сусідніх сегментах сигналу амплітуда тональної складової змінюється незначно. В даному алгоритмі порівнюються максимуми модулів спектрів потужності двох сусідніх сегментів сигналу S_{max}^{j} і S_{max}^{j+1} (де j — індекс сегмента) і обчислюється їх відносна різниця:

$$D_S = \frac{S_{max}^{j+1} - S_{max}^{j}}{S_{max}^{j}} \quad (2.7)$$

Порівняння величини D_s із заздалегідь обраним порогом T_d, дає шуканий результат: якщо $D_s < T_d$, то приймається рішення про наявність тональної складової в j-му фрагменті сигналу.

Кліпування – спотворення форми сигналу, що відбувається при перевантаженні підсилювача і при виході вихідного напруги підсилювача з його динамічного діапазону. На осциллограмі Кліпування зазвичай виглядає як обмеження сигналу по амплітуді. На слух кліппірованіе сприймається як поява зайвої дзвінкості, "металевого" звучання і може істотно знижувати якість обробки мови. Алгоритм детектування кліппірованія на основі аналізу гістограми сигналу наведено в роботі[8].

Наведені вище алгоритми детектування шумів застоссовані у бібліотеці Sphinx4.

2.3 Опис алогиртму веріфікації диктора

Зараз використовуються різні алгоритми створення таких систем. Наприклад, алгоритм динамічного перетворення часу (Dynamic Time Warping) використовується для текстозалежних систем. Цей алгоритм використовується для розпізнавання мовлення в тому випадку, коли дві різні людини сказали якусь одну фразу і треба

дізнатися, хто саме. Для цього комп'ютер порівнює дві «карти» голоси, відображені на синусоїдних графіках . Для порівняння досить всього два-три слова.

На рисунку 2.9 зображений синусоїдний графік порівняння двух голосів.

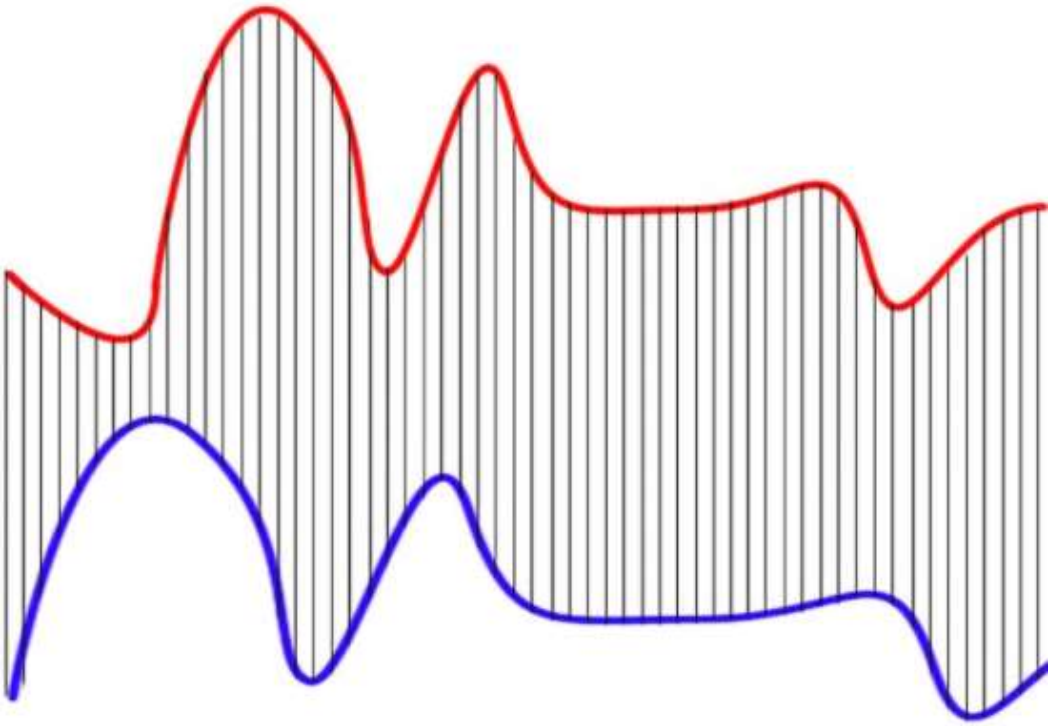

Рисунок 2.9 – Синусоїдний графік порівняння двух голосів

Але буває так, що ці графіки у людей дуже схожі. Щоб порівняти такі голоси, комп'ютер буде «деформувати» вісь часу одного або обох графіків, щоб досягти кращого вирівнювання. Помилки виникають через те, що порівнювані послідовності мають різну довжину і точка одного ряду буде розташована трохи вище або трохи нижче відповідної точки в іншому ряду. Тому алгоритму важко знайти видиме вирівнювання двох рядків. Проте він добре справляється з розпізнаванням окремих слів в обмеженому словнику. Це простий і відкритий для поліпшення алгоритм, відповідний для додатків в телефонах, автомобільних комп'ютерах або системах безпеки.

На рисунку 2.10 зображений синусоїдний графік з деформованною вісью часу.

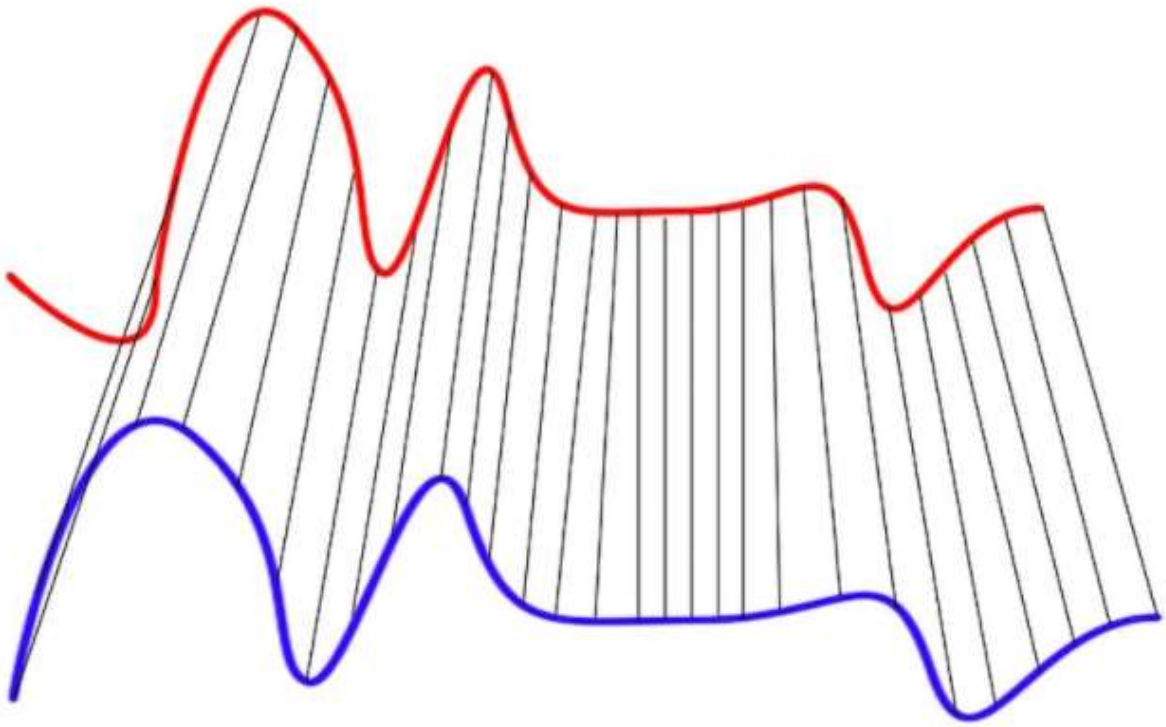

Рисунок 2.10 – Синусоїдний графік з деформованною вісью часу

Взаємодія людини і побутової техніки набагато розшириться, якщо керувати машиною звичайним голосом в реальному часі. Таку мету можна досягти, вирішуючи таку задачу: необхідно розпізнати певний ключове слово в природному потоці мовлення одиночного диктора. При наявності ключового слова сформувати условноооповестітельний сигнал, який в подальшому застосуванні може стати причиною вироблення якогось дії. Ключове слово може бути записано заздалегідь в словник. Непоганих успіхів у вирішенні подібних завдань домоглася корпорація Google, яка пропонує мовне введення при здійсненні пошуку в мережі Інтернет, але подробиці технології невідомі, тому є необхідність продовжити розробку технології розпізнавання мовлення. За класифікацією теорії розпізнавання сформульована задача відноситься до типу дікторозавісімого розпізнавання з обмеженим словником.

Для розпізнавання мовлення акустичний (мовної) сигнал за допомогою сприймають (мікрофона) і оцифровують (діскретізіруется) пристроїв і машинної обробки фіксується і перетворюється в цифрову форму. В результаті дискретизації безперервний (аналоговий) сигнал переводиться в послідовність чисел. Для роботи була прийнята частота дискретизації Fd = 8000 Гц і глибина кодування по рівню 16

біт двійкового коду. Це відповідає формату WAV, найбільш поширеній формату без стиснення, що застосовується в телекомунікаційних системах.

Перший етап розпізнавання мовлення – виділення різних ознак. Потім за допомогою деякої стратегії навчання формуються шаблони, з якими надалі будуть порівнюватися невідомі ділянки мовного сигналу. Традиційні моделі будови слухового апарату людини припускають наявність здатності аналізувати енергію звуку залежно від частотного інтервалу.

Математично таку гіпотезу зручно описувати апаратом субполосних матриць. Визначення енергії в заданому частотному інтервалі здійснюється наступним методом. Частотна вісь, нормована по відношенню до частоті дискретизації, розбивається на R рівновеликих інтервалу:

$$v_0 = 0,\ v_0 = \pi, v_r = \frac{r\pi}{R}, r = 1 \dots (R-1), \tag{2.8}$$

де ν, r – верхня межа r-го частотного інтервалу; R – кількість рівновеликих інтервалів, на які розбивається вісь частот.

Енергія в кожному частотному інтервалі оцінюється на основі виразу:

$$P_r = \vec{x}^T \cdot A_r \cdot \vec{x}, \tag{2.9}$$

де x – аналізований відрізок сигналу, тривалістю N відліків; T – знак транспонування; A_r – субполосная матриця з елементами виду:

$$A_r(i,k) = \begin{cases} \frac{\sin(v_r(i-k)) - \sin(v_{r-1}(i-k))}{\pi(i-k)}, & \text{при } i \neq k \\ \frac{(v_r - v_{r-1})}{\pi}, & \text{при } i \neq k \end{cases}, \tag{2.10}$$

де v_r – верхня межа r-го частотного інтервалу;

$$\begin{aligned} i &= 1 \dots N \\ k &= 1 \dots N \end{aligned} \tag{2.11}$$

Субполосні матриці мають всього декілька власних чисел відмінних від нуля. Така обставина прискорює обчислення квадратичної форми шляхом використання формули:

$$P_r = \sum_{k=1}^{J} \lambda_k^r \, (\vec{x}\vec{\mathrm{Q}}_k^r)^2 \,, \tag{2.12}$$

де λ_k^r – власні числа матриці A_r їх кількість J, розташовані по зростанню власні числа з індексом більше J незначні $\lambda_{J+k}^r \approx 0$, для практичних обчислень J = 8; $\left(\vec{x}\vec{\mathrm{Q}}_k^r\right)$ – скалярний добуток вектора аналізованого сигналу на власний вектор $\vec{\mathrm{Q}}_k^r$ матриці A_r, що відповідає власному числу λ_k^r.

Також треба зауважити, що формула (2.12) відповідає нейронної мережі з двома шарами. На рис. 2.11 оцінка енергії в r-му інтервалу частот виходить на виході другого шару, вагові коефіцієнти $\lambda_k^r, k = 1 \ldots K$ якого є власні числа від першого до K-го субполосной матриці Ar. Між першим шаром і другим діє функція активації, яка відповідає дії піднесення до квадрат. Перший шар складається з нейронів, у яких вагові коефіцієнти $\mathrm{q}_{ki}^r, k = 1 \ldots K, i = 1 \ldots N$ – це елементи власних векторів $\vec{\mathrm{Q}}_k^r$ матриці A_r.

Отримані таким чином оцінки енергії в частотних інтервалах представляють собою повну систему ознак в сенсі аддитивности.

$$\|\vec{x}\|^2 = \sum_{r=1}^{R} \vec{x}^T \cdot A_r \cdot \vec{x} \,, \tag{2.13}$$

Треба вибрати число відліків N для аналізу вихідного сигналу. Довжина вікна аналізу не повинна бути занадто великою, інакше будуть проявлятися ефекти накладення фонем один на одного. Велика довжина вікна аналізу також збільшує обсяг основних операцій (складань і умножений), необхідних для подальших перетворень вихідного сигналу.

Маленька довжина вікна аналізу погіршує частотне дозвіл досліджуваного сигналу. З цих умов довжину вікна аналізу N треба вибирати з діапазону від 60 до 160 відліків. На рисунку 2.11 зображена нейронна мережа для оцінки енергії в межах інтервалу частот.

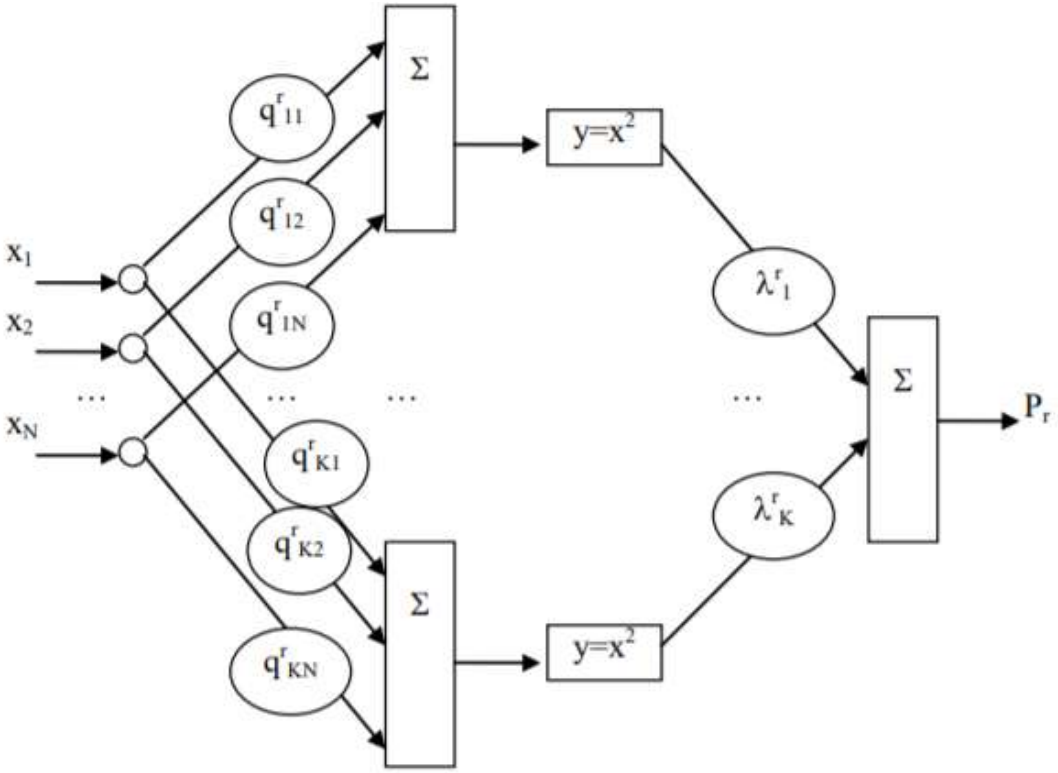

Рисунок 2.11 – Нейронна мережа для оцінки енергії в межах інтервалу частот

Для проведення експерименту взяті субполосние матриці зі значеннями N = 60 і R = 15. Кількість відмінних від нуля власних чисел взято J = 8. З звукового файлу, містить мовний сигнал, надходили відліки n x, що формують вхідний сигнал $\vec{x}$

$$\vec{x}_n = [x_{n-N+1} \quad x_{n-N+2} \dots x_{n-1}\, x_n]^T, n > N, \tag{2.14}$$

де n – це номер відліку, дискретне час.

Видається раціональним проводити дослідження сигналу, зрушуючи вікно аналізу на один відлік. Тоді значення оцінок енергій можна записувати з таким же позначенням моменту часу, як і у вхідного сигналу.

$$P_r = \sum_{k=1}^{J} \lambda_k^r \, (\vec{x}\vec{Q}_k^r)^2 \, , r = 1 \dots R \, , \qquad (2.15)$$

Оцінки енергій доцільно усереднити на деякому відрізку часу [n - Ts, n], n> Ts.

$$\beta_r(n) = \frac{1}{Ts} \sum_{k=n-Ts}^{n} P_r(k) \; , r = 1 \dots R \, , \qquad (2.16)$$

Величина інтервалу Ts повинна бути сумірною з довжиною ділянки мовного сигналу, на якому він стаціонарен. Виходячи з практичних спостережень, кількість фонем в одну секунду часу не перевищує 25. При частоті дискретизації 8000Гц на одну фонему доводиться в середньому 320 відліків. З урахуванням середньоквадратичних відхилень від середнього однієї фонеми, Ts раціонально вибрати з діапазону від 100 до 200 відліків, т. к. точність усереднення падає зі зменшенням числа членів. Усереднення за способом «Ковзне середнє» являє собою CIC-фільтр, дуже економний по обчислювальної складності.

$$\beta_r(n) = \frac{1}{Ts} \big(P_r(n) + \beta_r(n-1) - \; P_r(n-Ts) \big), r = 1 \dots R \, . \qquad (2.17)$$

В експерименті було взято Ts = 200.

Після усереднення стає можливим аналізувати оцінки енергії не для всіх відліків, а тільки через кожні Ts / 2 відліків.

$$\beta 1_r(n1) = \beta_r(n), n = \frac{Ts}{2} n1 \, , r = 1 \dots R \, , n1 = 1,2,3 \qquad (2.18)$$

Для проведення найпростішого етапу навчання в програмі Simulink була складена комп'ютерна модель. На вхід моделі було подано звуковий файл, в якому містилося кілька разів вимовлене слово «п'ять». Результат обробки за формулами (2.14) – (2.18) представлений на рис. 2.12.

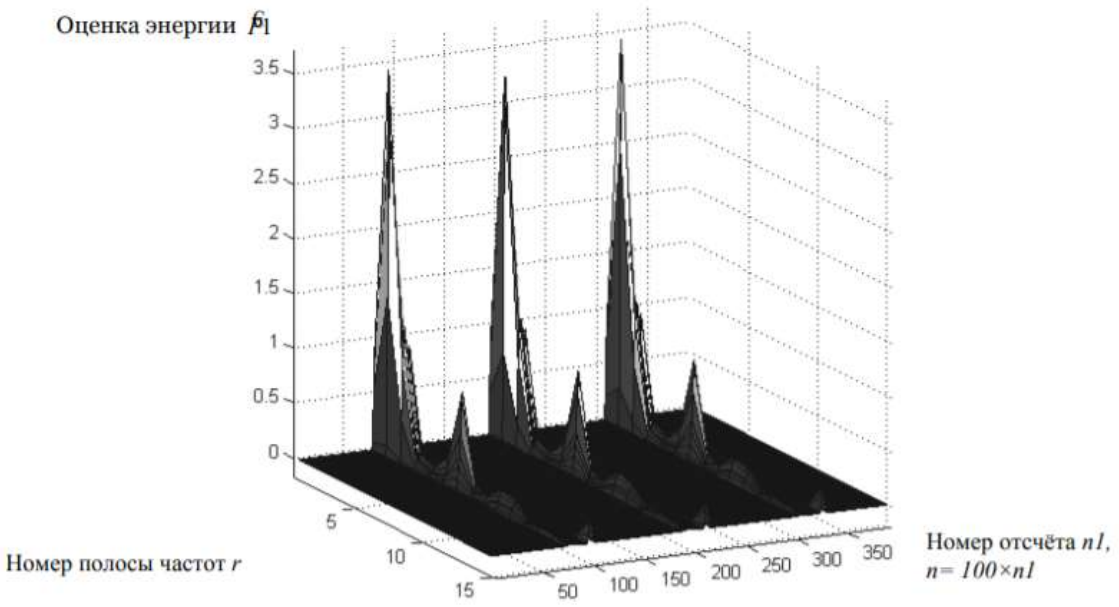

Рисунок 2.12 – Оцінка енергій для слова «п'ять» після проведення експерименту

Вручну виділено всього слова, рисунку 2.13, і збережено для подальшого використання в якості шаблону. Шаблон являє собою матрицю Pt з кількістю рядків рівній кількості аналізованих частотних інтервалів. кількість стовпців відповідає кількості відібраних відліків N_{Pt} .

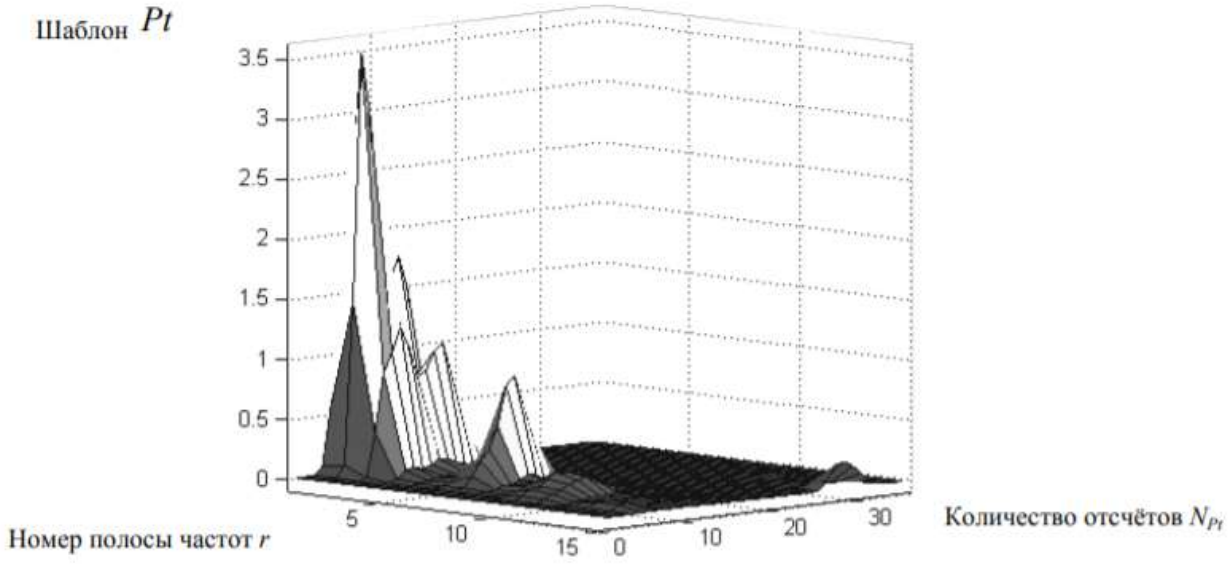

Рисунок 2.13 – Шаблон для слова «п'ять»

Для вирішення поставленого завдання розпізнавання далі треба порівнювати відрізок мовного сигналу $\beta_1\,(n1), n1 \in\ [n1b, n1e]$, де $n1b$ – початок відрізка, $n1e$ – кінець відрізка, з записаним в словник шаблоном Pt. Основна складність порівняння – мінливість слова, як по тривалості, так і за звучанням. Таким чином, треба порівняти дві матриці між собою, кількість рядків в матрицях однакове і відповідає кількості розглянутих частотних інтервалів R, а кількість стовпців – різний. У матриці шаблону кількість стовпців фіксоване N_{Pt} , у матриці розпізнається мовного сигналу кількість стовпців $N1 = n1e - n1b$ береться за тривалістю звучання передбачуваного слова. При такому підході ці матриці являють собою послідовності з різною кількістю елементів, якими є стовпці [17].

Для вимірювання схожості двох послідовностей, які мають різний кількість елементів, часто застосовується алгоритм динамічного спотворення (dynamic time warping). Порівняння проводиться за правилом «найкращого відповідності». За цим алгоритмом спочатку треба скласти матрицю D з відстанями між елементами послідовностей. Пропонується використовувати відносні заходи відмінності при визначенні відстані:

$$D(i,k) = \frac{\sum_{r=1}^{R}|\beta 1(r,i) - Pt(r,k)|}{\sum_{r=1}^{R}|\beta 1(r,i) + Pt(r,k)|}, i = 1 \dots N1, k = 1 \dots\ N_{Pt}\,, \qquad (2.19)$$

Потім складається матриця *C* з накопичувальної дистанцією між поодинокими елементами двох послідовностей:

$$C(i,k) = D(i,l) + \min\big(C(i,k-1), C(i-1,k), C(i-1,k-1)\big), \qquad (2.20)$$

де мінімум береться серед трьох сусідніх елементів;

$$C(1,1) = 0, C(1,k) = \infty, C(i,1) = \ \infty; \quad k = 2 \dots\ N_{Pt} \quad i = 2 \dots N1. \qquad (2.21)$$

У розв'язуваної задачі відліком є стовпець зі значеннями оцінок енергії. Потім в матриці С відновлюється шлях з точки ($N1$, N_{Pt}) в точку (1,1) по правилом руху в бік з найменшим значенням:

$$(i_s, k_s) = arg\ \min\big(C(i, k-1), C(i-1, k), C(i-1, k-1)\big), \quad (2.22)$$

де s – це номер кроку по шляху від (N1, NPt) до (1,1). Підсумкове кількість кроків S НЕ постійно, кроки здійснюються поки обидва індекси i та k не зміняться від своїх максимальних значень $N1$ і N_{Pt} до 1.

Потім обчислюється загальна відмінність SH між двома послідовностями. У класичному варіанті алгоритму:

$$SH = \sum_{s=1}^{S} D(i_s, k_s), \quad (2.23)$$

в даній роботі пропонується:

$$SH = \max_{s}\big(D(i_s, k_s)\big), s = 1 \dots S. \quad (2.24)$$

При порівнянні невідомої послідовності з шаблоном рішення про відповідність шаблоном приймається після порівняння SH із заданим порогом.

Для експерименту один диктор надиктував п'ять звукових файлів з однаковим змістом: вимовлені по порядку числівники: «1», «2», «3», «4», «5», «6», «7», «8», «9» і «0». Мовний сигнал з звукових файлів порівнювався з заздалегідь збережених шаблоном для слова «п'ять». Оцінки енергій $\beta 1(n1)$ n на кроці n1формувалісь в матрицю з кількістю стовпців N1 = 50, т. к. у шаблону кількість стовпців виявилося N_{Pt} = 36.

Матриця невідомого сигналу береться на більшу кількість відліків, ніж у шаблону (50> 36), в розрахунку на можливе повільне проголошення шуканого слова.

Результат моделювання на рис. 2.14 для мовного сигналу з файлу №2 і на рис. 2.15 – з файлу №4.

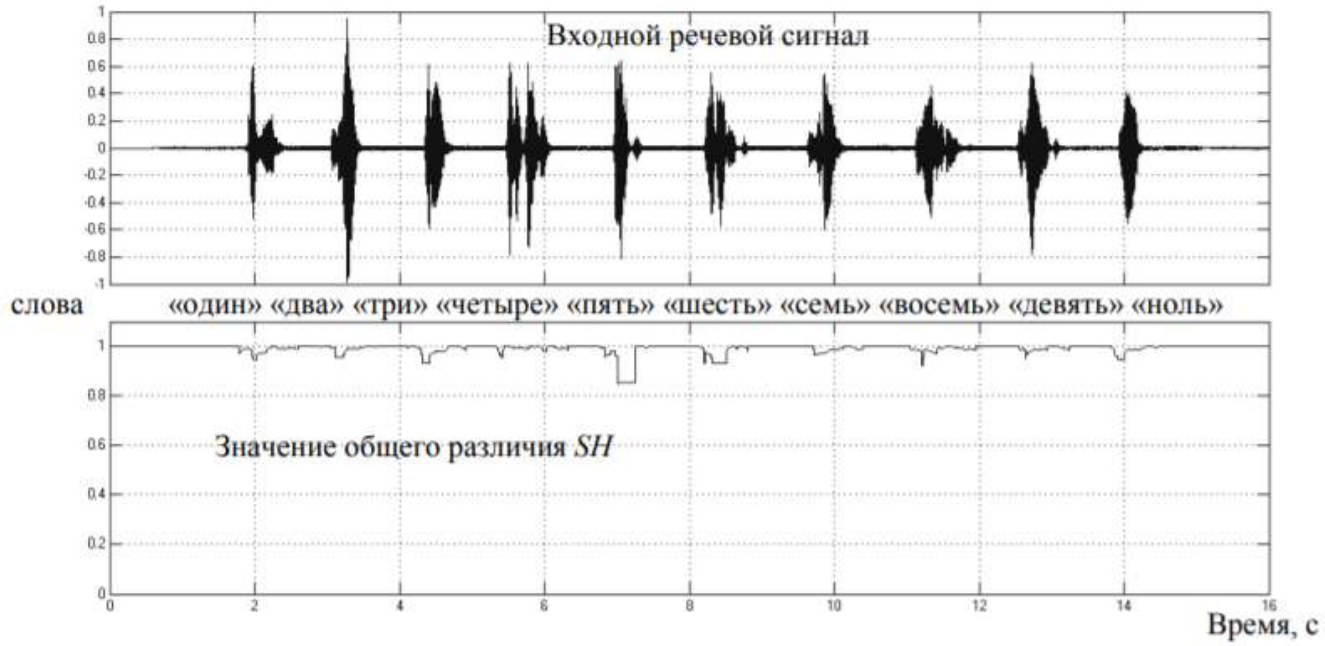

Рисунок 2.14 – Графіки вхідного сигналу і величини відмінності від шаблону після використання відносної міри. Вхідний сигнал з файлу №2

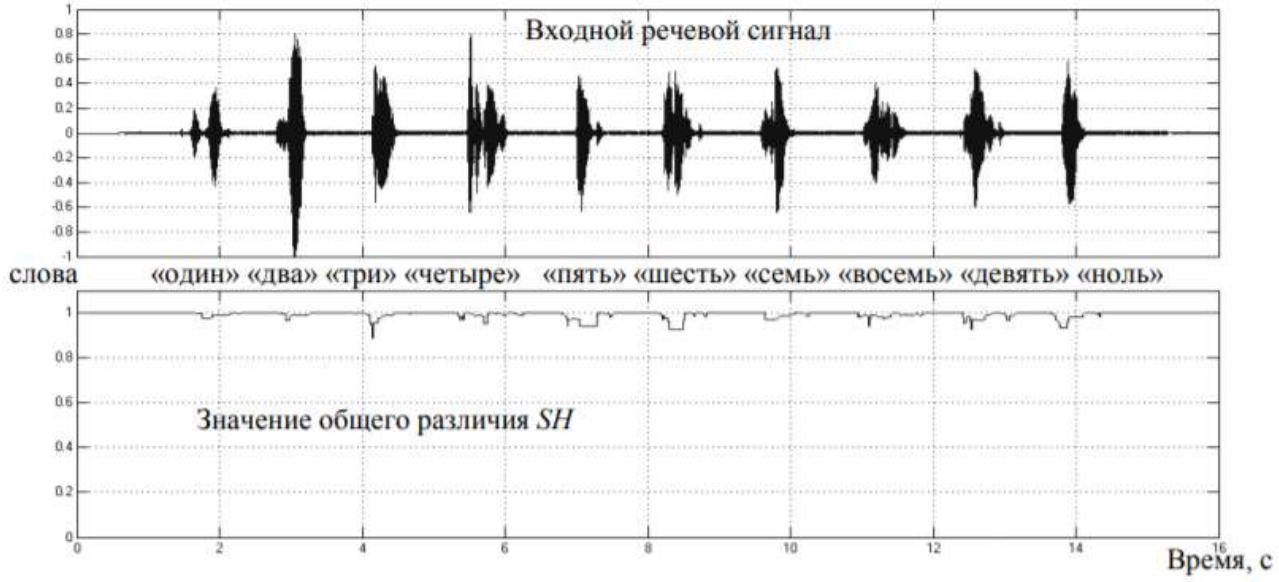

Рисунок 2.15 – Графіки вхідного сигналу і величини відмінності від шаблону після використання відносної міри. Вхідний сигнал з файлу №4

На рис. 2.14 графіки показують чітку реакцію, а на рис. 2.15 більш переконливою здається реакція на слово «шість», хоча для слова «п'ять» відміну від шаблону менше, ніж для інших слів.

Для підвищення чіткості при аналізі відмінності між шаблоном і вхідним сигналом можна скористатися нелінійним перетворенням величини оцінок енергії. Згідно емпіричному психофізіологічного закону Вебера-Фехнера сила відчуття пропорційна логарифму інтенсивності подразника. Інший вчений, Стівенсон, запропонував використовувати ступеневу функцію для опису залежності сили відчуттів від величини роздратування.

Показник статечної функції для різних відчуттів використовується різний від 3,5 до 0,67 в експериментах Стівенсона.

Для вирішення поставленого завдання розпізнавання можна застосувати нелінійність витяг квадратного кореня з величини оцінок енергій, що буде схоже на статечної закон Стівенсона, але не внесе помітних обчислювальних складнощів.

Нові оцінки з урахуванням нелінійностей:

$$W(n1) = \sqrt{\beta 1(n1)},$$
$$W_{Pt} = \sqrt{Pt}. \qquad (2.25)$$

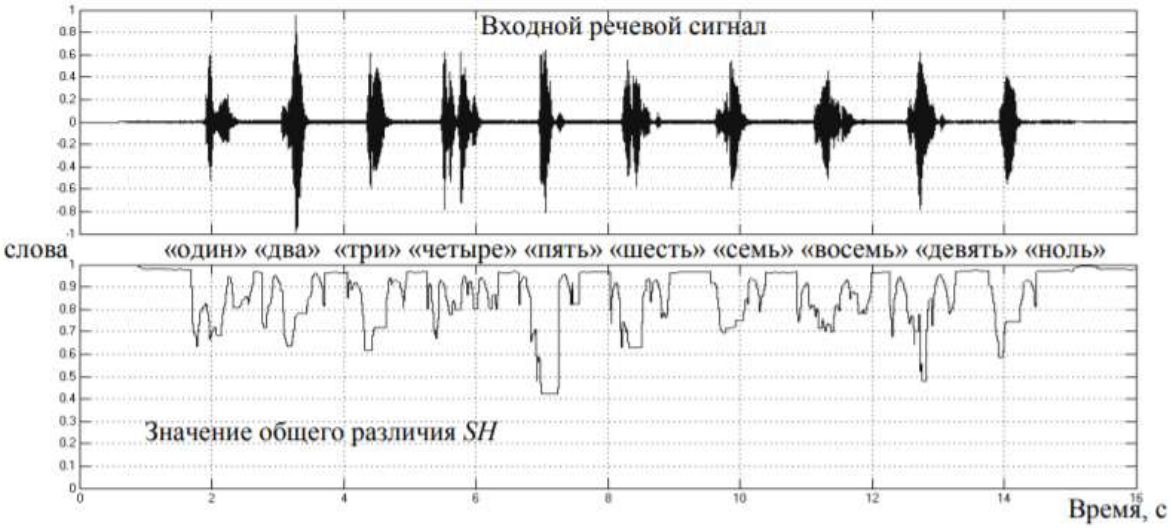

Рисунок 2.16 – Графіки вхідного сигналу і величини відмінності від шаблону після використання відносної міри і нелінійності. Вхідний сигнал з файлу №2

Результат моделювання при наявності нелінійності на рис. 2.17 і 2.18.

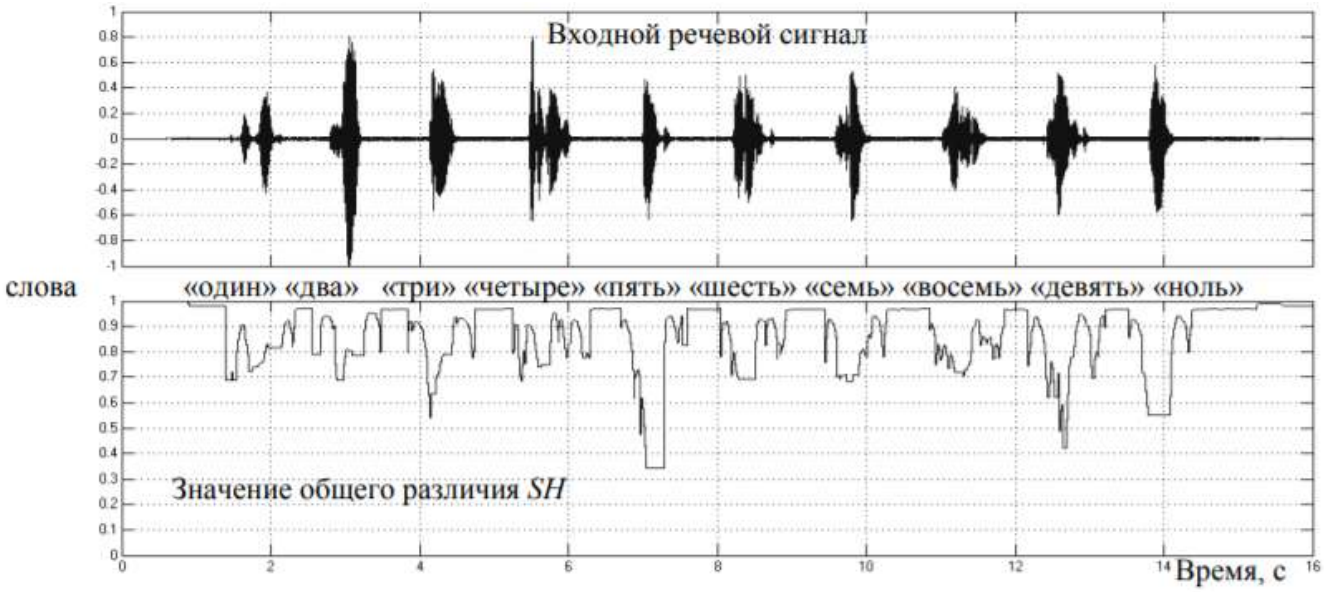

Рисунок 2.17 – Графіки вхідного сигналу і величини відмінності від шаблону після використання відносної міри і нелінійності. Вхідний сигнал з файлу №4

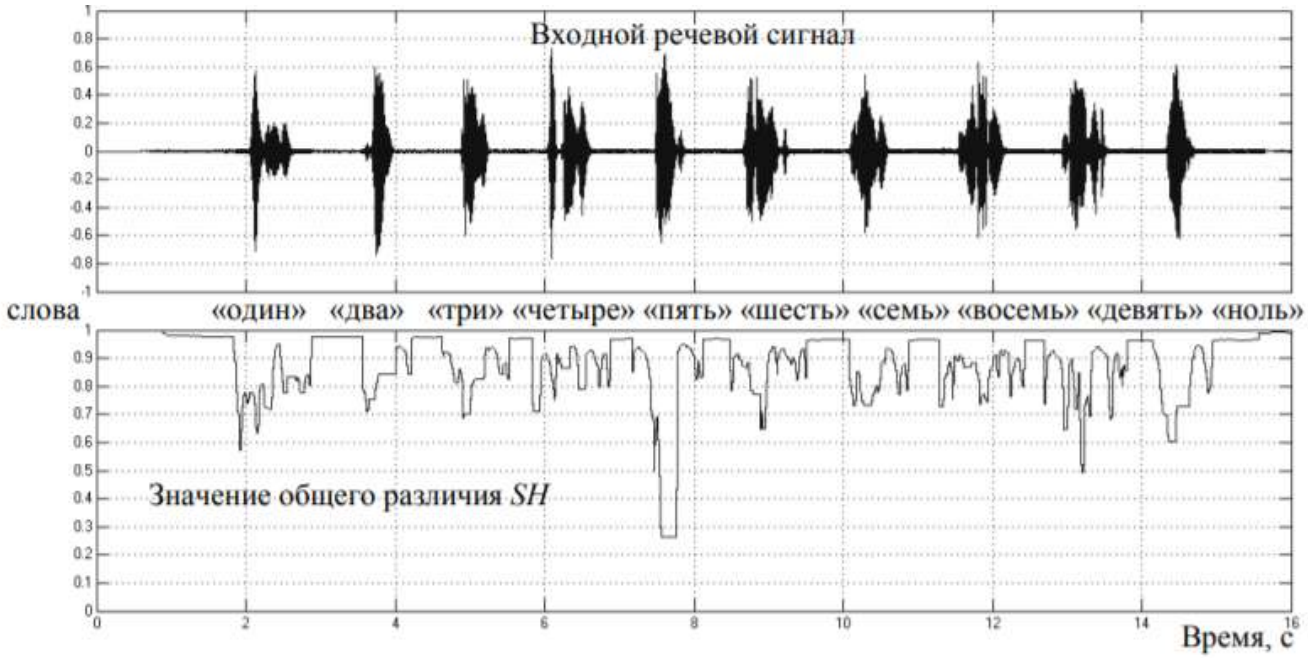

Рисунок 2.18 – Графіки вхідного сигналу і величини відмінності від шаблону після використання відносної міри і нелінійності. Вхідний сигнал з файлу №1

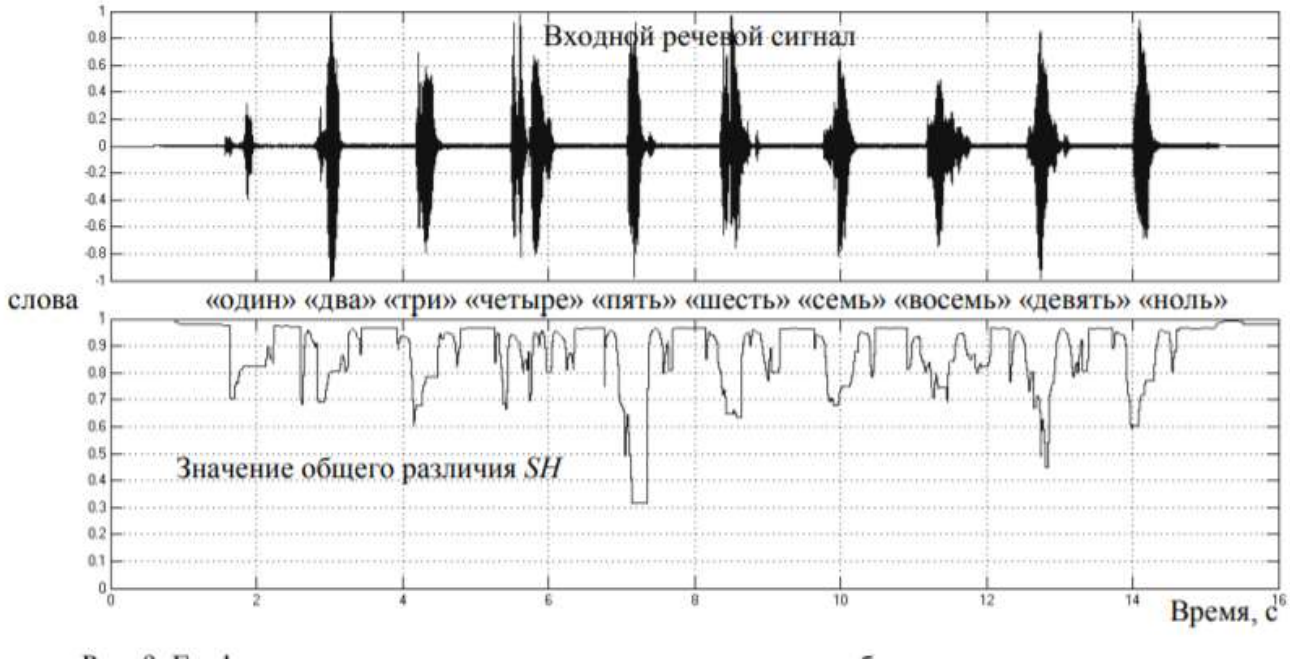

Рисунок 2.19 – Графіки вхідного сигналу і величини відмінності від шаблону після використання відносної міри і нелінійності. Вхідний сигнал з файлу №3

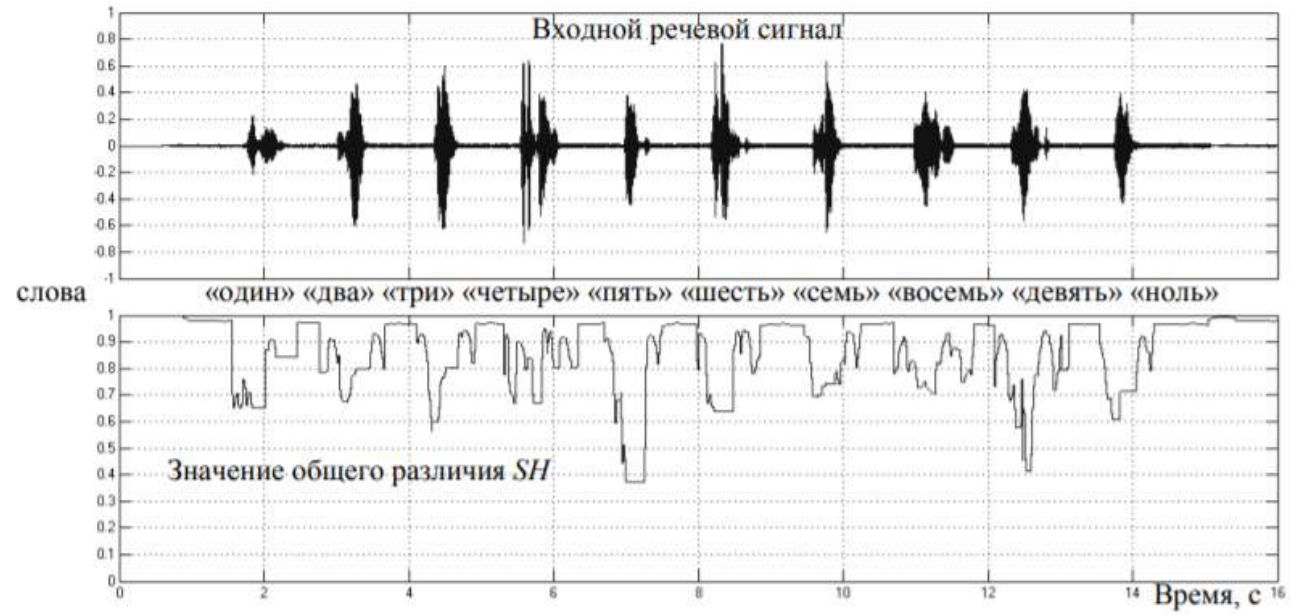

Рисунок 2.20 – Графіки вхідного сигналу і величини відмінності від шаблону після використання відносної міри і нелінійності. Вхідний сигнал з файлу №5

Застосування нелінійності дозволило підвищити розрізнювальну здатність алгоритму. Мінімум для слова «п'ять» виділяється сильніше по відношенню до інших

слів і графіками попередніх експериментів. Для наочності на рис. 2.18-2.20 наведені графіки для мовних сигналів з файлів №1, №3, №5. Поведінка графіків однакове.

Граничне значення необхідно визначити з урахуванням різноманітності вимови ключового слова. Для цього з наявних звукових файлів вручну були виділені і збережені в змінні матриці $\beta 1_i, i = 1,2,3,4,5$ з оцінками енергії для слова «п'ять». Потім матриці порівнювалися кожна з кожною алгоритмом динамічного спотворення з застосуванням відносної міри між стовпцями і нелінійним перетворенням оцінок енергії шляхом вилучення квадратного кореня.

$$SH_{ij} = DTW(\sqrt{\beta 1_i}, \sqrt{\beta 1_j}, \tag{2.26}$$

де i, j = 1,2,3,4,5 – індекси файлів, звідки було взято слово «п'ять», $\beta 1_i$, € $\beta 1_j$ – матриці ознак для слова «п'ять», елементи матриці обчислюються за формулою (4) з подальшим застосуванням змінного середнього (2.17); кількість рядків у матриці дорівнює кількості частотних інтервалів, кількість стовпців відповідає тривалості звучання слова після проріджування (2.18), визначено вручну; $DTW()$ – функція алгоритму динамічного перекручування, описана формулами (2.19) – (2.22), (2.24).

На відміну від безперервного мовного сигналу, немає необхідності ковзати вікном аналізу, відзначаючи величину відмінності в кожен момент часу. На даному етапі є фіксовані матриці, що описують слово «п'ять». Тому в результаті порівняння виходить одне число, а не графік в часі. Результати парних порівнянь показують, наскільки різноманітно вимовлялося слово «п'ять» одним диктором.

3 ПРОЕКТУВАННЯ І РОЗРОБКА ПРОГРАМНОГО ЗАСОБУ

Програмний засіб, що розробляється в рамках атестаційної роботи – програмний засіб для тестування методу аутентифікації диктора на основі алогритму динамічного перетворення часу.

Програмний засіб повинен мати логічно декомпозовану структуру. Користування програмою має бути простим.

Для розробки програмного засобу розпізнавання та виконання голосових команд треба скласти проект даного програмного засобу. Також необхідно підібрати апаратне і програмне забезпечення, яке задовольнить потребам розробки програмного засобу при мінімальних затратах.

3.1 Створення проекту системи аутентифікації голосових команд

Для проектування на основі предметної області була використана методологія IDEF0 для створення функціональної декомпозиції процесів.

IDEF0 (Function Modeling) – методологія функціонального моделювання і графічного опису процесів, призначена для формалізації і опису бізнес-процесів. Особливістю IDEF0 являється її акцент на ієрархічному представленні об’єктів, це спрощує розуміння предметної області. В IDEF0 розглядаються логічні зв’язки між роботами, відображаються сигнали управління. Опис виглядає як «чорний ящик» з входами, виходами, управлінням і механізмом, який поступово деталізується до необхідного рівня. Така модель – одна з найбільш прогресивних моделей, вона використовується в організації багатьох проектів.

DFD (data flow diagram) – діаграми потоків даних. Так називається методологія графічного структурного аналізу, що описує зовнішні по відношенню до системи джерела і адресати даних, логічні функції, потоки даних і сховища даних, до яких здійснюється доступ. Ддин з основних інструментів структурного аналізу і проектування інформаційних систем, що існували до широкого поширення UML [10].

На рисунках 3.1-3.2 представлені контекстна діаграма та діаграми декомпозиції процесів.

Рисунок 3.1 – Контекстна діаграма

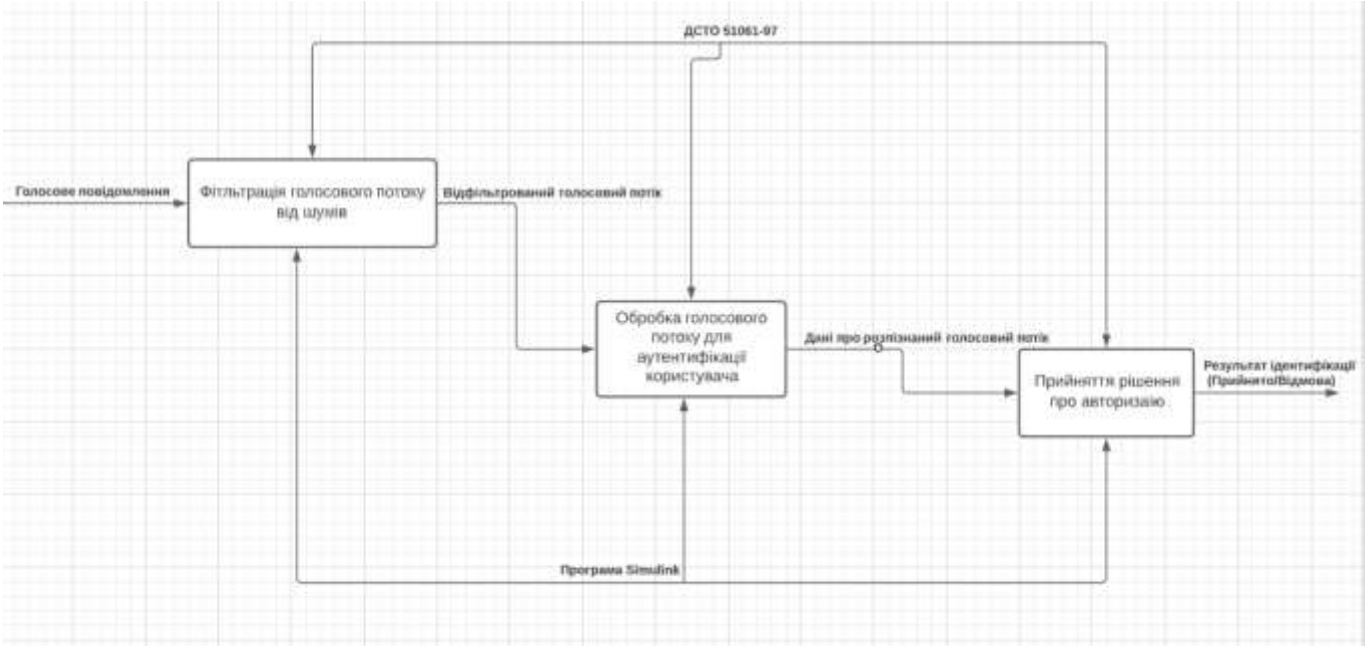

Рисунок 3.2 – Декомпозиція основного бізнес-процесу

На рисунку 3.3 представлена DFD діаграма.

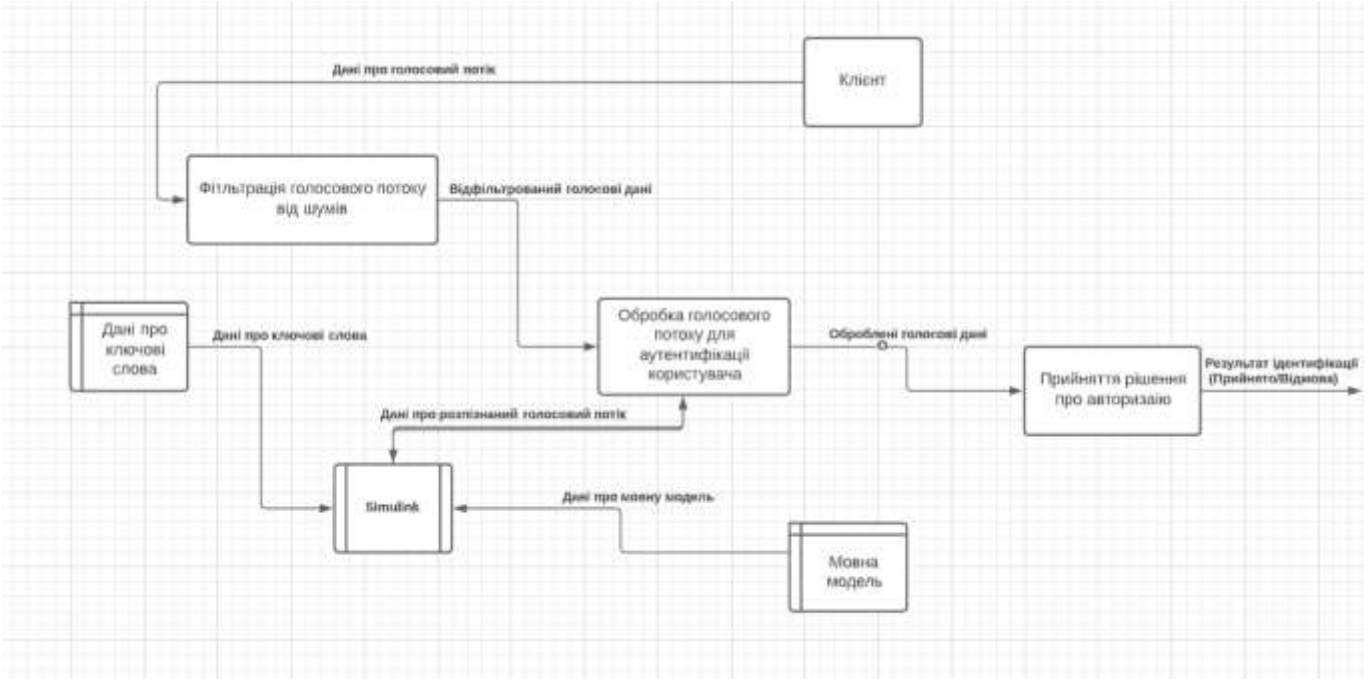

Рисунок 3.3 – DFD діаграма

UML – мова графічного опису для об'єктного моделювання в області розробки програмного забезпечення, для моделювання бізнес-процесів, системного проектування та відображення організаційних структур.

UML є мовою широкого профілю, це – відкритий стандарт, який використовує графічні позначення для створення абстрактної моделі системи, званої UML-моделлю. UML був створений для визначення, візуалізації, проектування та документування, в основному, програмних систем. UML не є мовою програмування, але на підставі UML-моделей можлива генерація коду.

Для того, щоб описати, як буде використовуватись розроблюваний програмний засіб на концептуальному рівні слід скласти діаграму прецедентів.

Прецедент – це функціональність системи, яка дозволяє користувачеві отримати деякий значущий для нього, відчутний і такий, що можна виміряти, результат. Кожен прецедент відповідає окремому сервісу, що надається системою, яка моделюється, у відповідь на запит користувача, тобто визначає спосіб використання цієї системи. Варіанти використання найчастіше застосовуються для специфікації зовнішніх вимог до проектованої системи або для специфікації функціональної поведінки вже існуючої системи. Крім цього, варіанти використання

неявно описують типові способи взаємодії користувача з системою, що дозволяють коректно працювати з сервісами, які надає система [11].

Розроблюваний програмний засіб призначений для написання і відправки листів за допомогую голосу. З фреймворком буде взаємодіяти єдина діюча особа.

Діаграма варіантів використання системи, представлена на рисунку 3.4, була складена на підставі вимог до програмного засобу.

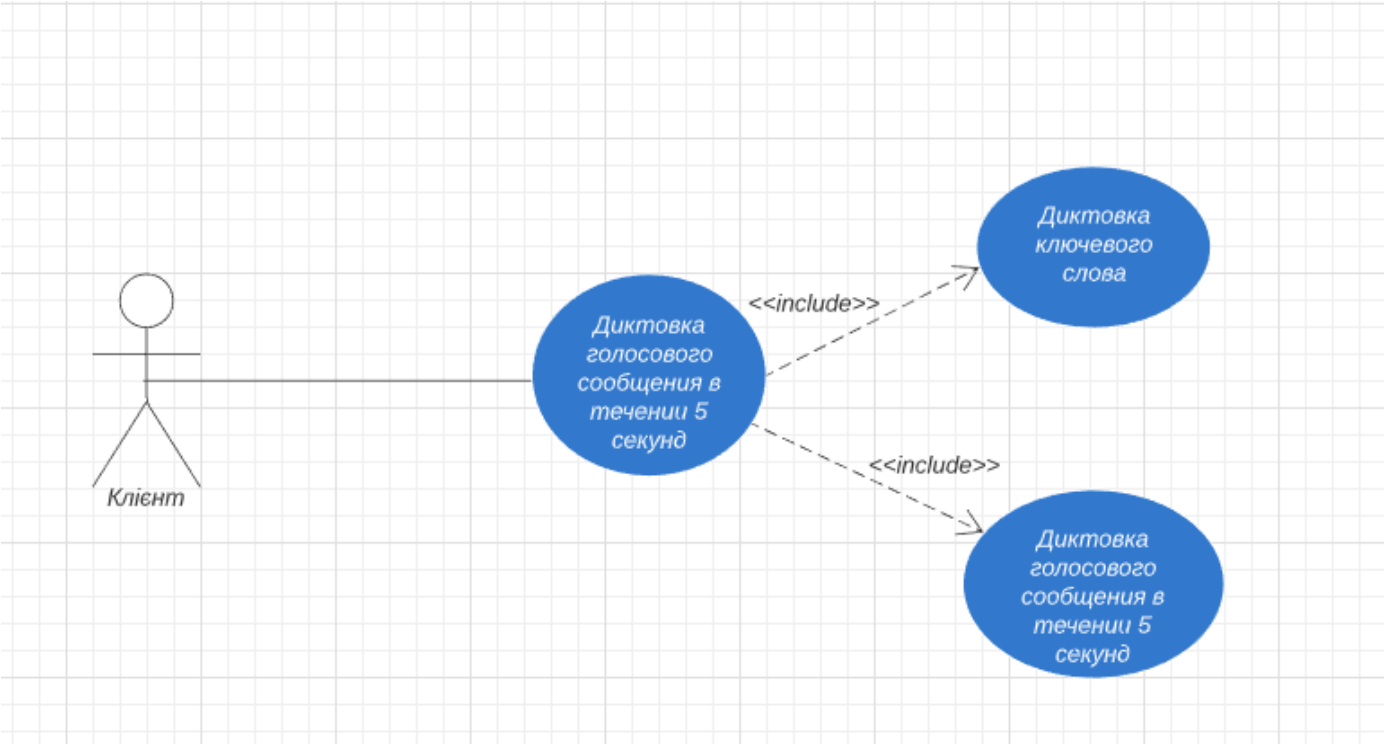

Рисунок 3.4 – Діаграма варіантів використання системи

Для написання програми слід створити діаграмау класів яка допоможе визначитися з програмною архітектурою.

Діаграма класів – структурна діаграма мови моделювання UML, що демонструє загальну структуру ієрархії класів системи, їх кооперацій, атрибутів (полів), методів, інтерфейсів і взаємозв'язків між ними.

На рисунку 3.5 зображена діаграма класів, яка була використана під час написання коду програмного засобу.

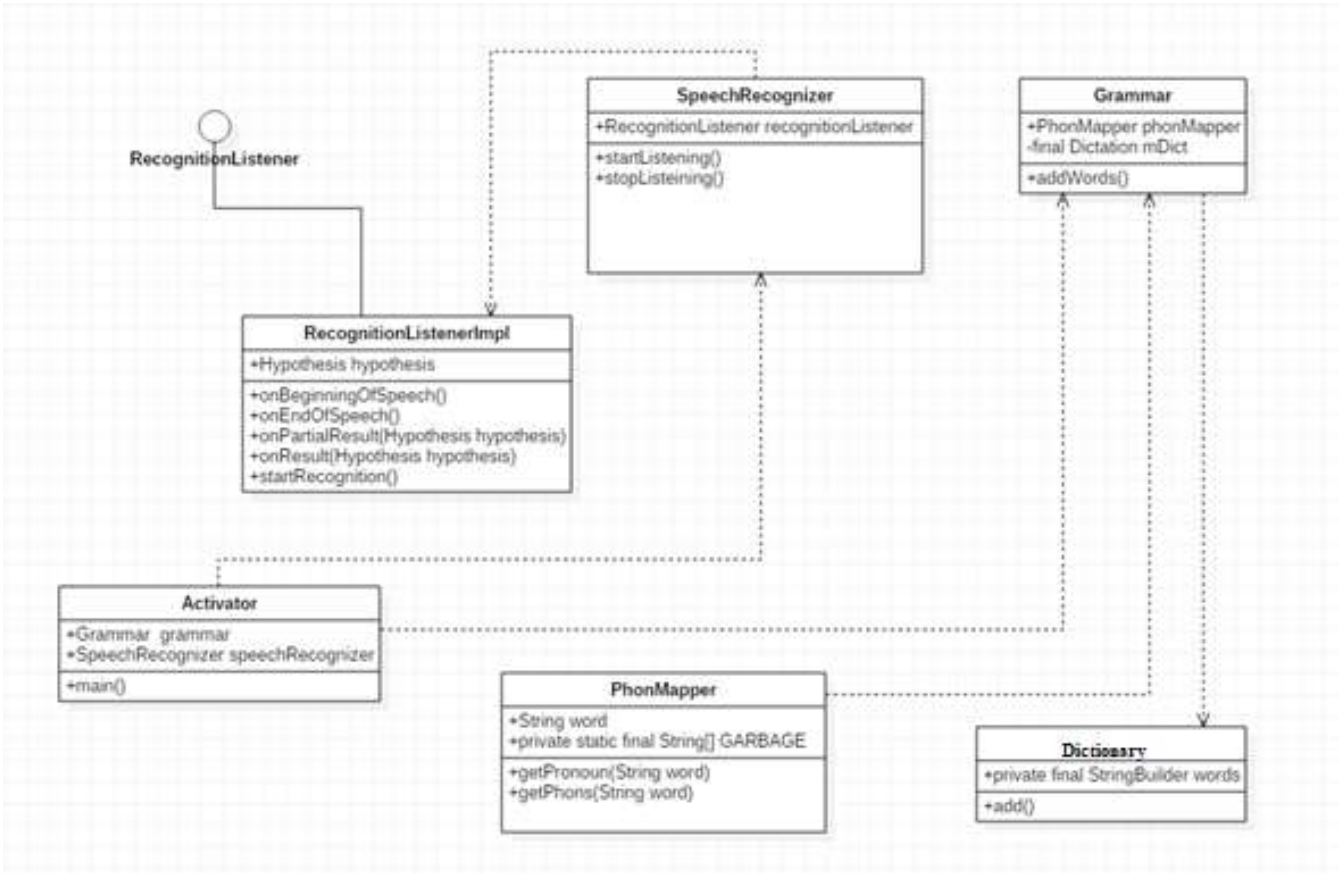

Рисунок 3.5 – Діаграма класів

3.2 Вибір і обґрунтування технічного забезпечення

Для забезпечення необхідного рівня продуктивності слід використовувати комп’ютер з тактовою частотою процесора не менше 2 Ггц і обсягом оперативної пам’яті не менше 4 Гб. При даних значеннях основних характеристик процесора компіляція коду і запуск тесту буде відбуватися без значних затримок. В роботі середовища розробки не повинно бути збоїв.

Програма буде працювати за допомогою гнучкої бібліотеки розпізнавання мовлення Simulink.

Simulink – середа динамічного междисциплінарного моделювання сложних технічних систем та основний інструмент для модельно-орієнтованого проектування. Його основним інтерфейсом є графічний інструмент для побудови діаграми та настраїваного набору бібліотечних блоків. Він пропонує тестову інтеграцію з останньою середою MATLAB і може використовувати MATLAB, або створити

сценарії з нього. Simulink широко використовується в автоматичному управлінні та цифровій обробці сигналів для багатоденного моделювання та проектування на основі моделей.

У співпраці з іншими своїми продуктами Simulink може автоматично генерувати вихідний код на мові C для реалізації системи в режимі реального часу. Після ефективності та гнучкості коду покращуються, він стає все більш широко розповсюдженим для виробничої системи, у доповненні до того, що є інструментом для розробки вбудованої системи із-за його гнучкості та можливостей для побудови ітерацій, Вбудований кодер створює код, достатньо ефективний для використання в вбудованих систем [12].

3.3 Вибір и обґрунтування мовних і програмних засобів

Програма бути розвроблятися на мові програмування Java. До переваг цієї мови можна віднести обчислювальну продуктивність. Мова спроектована так, щоб дати програмісту максимальний контроль над усіма аспектами структури і порядку виконання програми. Програма вона буде працювати з велким обсягом даних, тож підтримка продуктивності такого продукту важлива для швидкого відгуку системи.

Враховуючи основні характеристики та поширеність використання мов програмування у сфері автоматизованого тестування, вдалим вибором є мова програмування Java. Переважна більшість інструментів автоматизації тестування підтримує цю мову програмування. Серед основних характеристик Java можна виділити наступні: простий об’єктно-орієнтований синтаксис, незалежність від архітектури, висока продуктивність виконання.

Однією з найбільш зручних середовищ розробки на Java вважають Inteliji IDEA. Inteliji IDEA має механізми аналізу коду, який спрощує простежування зв’язаності коду, перехід до шуканої реалізації, знаходження помилок та здійснення рефакторингу.

Серед інструментів складання(будування) проекту був обраний Gradle – система автоматичного складання, побудована на принципах Apache Ant і Apache

Maven, але надає DSL на мовах Groovy і Kotlin замість традиційної XML-подібної форми подання конфігурації проекту.

На відміну від Apache Maven, заснованого на концепції життєвого циклу проекту, і Apache Ant, в якому порядок виконання завдань (targets) визначається відносинами залежності (depends-on), Gradle використовує спрямований ациклічний граф для визначення порядку виконання завдань.

Gradle був розроблений для розширюваних многопроектной збірок, і підтримує інкрементальні збірки, визначаючи, які компоненти дерева збірки не змінилися і які завдання, залежні від цих частин, не вимагають перезапуску [13].

Найважливіша задача у виборі програмних засобів для побудови тестівого фреймворку – вибір основного інструментального засобу автоматизованого тестування, оскільки від даного засобу залежить ефективність процесу перевірки програмного продукту. Слід підібрати оптимальний інструмент, який надає більшу частину необхідної функціональності та має мінімальні обмеження у важливих для проекту областях.

3.4 Опис інтеграційного тестування

Інтеграційне тестування – одна з фаз тестування програмного забезпечення, при якій окремі програмні модулі об'єднуються і тестуються в групі. Зазвичай інтеграційне тестування проводиться після модульного тестування і передує системному тестування [14]. Інтеграційне тестування в якості вхідних даних використовує модулі, над якими було проведено модульне тестування, групує їх в більш великі безлічі, виконує тести, певні в плані тестування для цих множин, і представляє їх в якості вихідних даних і вхідних для подальшого системного тестування. Метою інтеграційного тестування є перевірка відповідності проектованих одиниць функціональним, прийомним і вимогам надійності. Тестування цих проектованих одиниць – об'єднання, безлічі або групи модулів – виконується через їх інтерфейс, з використанням тестування «чорного ящика».

4 АНАЛІЗ РЕЗУЛЬТАТІВ, ОТРИМАНИХ ЗА ДОПОМОГОЮ ПРОГРАМНОГО ЗАСОБУ

Для тестування алгоритму ідентифікаціх диктора був розроблений програмний засіб, на вхід якого поступає голосовий сигнал, а на виході маємо результат ідентифікації – “Ідентифіковано/Відмова”. Розпізнавання мовлення виконується за допомогою утиліти Simulink.

Тестування проводилося на десяти користувачах. У програмі заздалегідь були сформовані десять мовних моделей користувачів яких потрібно ідентифікувати. Ключовим словом було обрано слово “п’ять”. Користувачі вимовляли цифри від одного до десяти, програма повинна була виділити це слово з голосового потоку, сформувати вхідну голосу модель і порівняти її з голосовими моделями в базі. Кожен користувач пробував авторизуватися по десять разів.

Для значення *FRR* необхідно зібрати статистику за N_{FRR-} – кількість відмов у аутентифікації користувача, $N_{FAR\ all}$ – загальна кількість спроб пройти аутентифікацію користувачем.

Для значення *FAR* необхідно зібрати статистику за N_{FAR+} – кількість успішних автентифікацій зловмисника, $N_{FAR\ all}$ – загальна кількість спроб пройти автентифікацію зловмисником. Результати приведені в таблиці 4.1.

Таблиця 4.1 – Результати тестування

Номер мнової моделі	Кількість успішних автентифікацій зловмисника	Кількість відмов у автентифікації користувача
1	4	2
2	5	3
3	3	4
4	3	4
5	4	2

Продовження таблицы 4.1

6	4	3
7	3	2
8	5	3
9	4	3
10	4	3

У середньому значення *FAR* = 3.9 та *FRR* = 2.9.

Таблиця 4.2 – Порівняння отриманих результатів з аналогами

Назва	*FAR*	*FRR*	*Кількість осіб*
Agnitio	2,1%	1,5%	24
Nuance	2,6%	1,8%	33
Voice Systems	1,1%	2,3%	14
VoiceTrust	2,5%	2,8%	25

Дослідження були присвячені оцінці якості аутентифікації диктора за методом динамічного перетворення часу. Було проведено ряд експериментів. Для цього за допомогою програми Simulink були створені мовні моделі, які порівнювалися з голосовими потоками поданими на вході.

Після тестування на десяти моделях був зроблений загальний висновок що при зростанні кількості успішник автонтифікацій зловмисника зменшується кількість відмов у автонтифікації користувача. Це послабляє рівень безпеки, однак спрощуж користування програмним засособом. Отже цей метод більш підходиь для автризаціх на мобільних пристроях чи персональних комп’ютерах, він є достатньо безпечний у повсякденні и коли користувач не хочу витрачати багато часу на авторизацію.

ВИСНОВКИ

Розпізнавання мовлення широко застосовується у сучасних інформаційних технологіях.

У ході виконання атестаційної роботи був проведений аналіз предметної області і був розроблений програмний засіб ідентифікації мовлення за допомгою ключового слова. Цільова аудиторія користувачів такого програмного засобу – системи розмежування доступу, при авторизації на персональному пристрої, при управлінні різними пристроями і т.д.

В результаті проведеного аналізу особливостей різних підходів до розробки програмних засобів для ідентифікації мовлення був зроблений висновок, що найбільш вигідно у цьому випадку використовувати метод розпізнавання ключового слова на основі субполосного перетворення з використанням алгоритму динамічного спотворення, тож він будв модифікованй для зростання ефективності.

Функціональність методу була перевірена за допомогою тестового програмного засобу та утиліті Simulink. Тесу вання проводилося на десяти користувачах. У програмі заздалегідь були сформовані десять мовних моделей користувачів і з їх використанням були виконані тести. Процес запуску набору тестів і їх результати були проаналізовані.

Результати пройшли опробацію на I Міжнародній науково-практичній конференції “PRIORITY DIRECTIONS OF SCIENCE AND TECHNOLOGY DEVELOPMENT” 27 - 29 вересня.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Тейлор, Р. Шум / Р. Тейлор; пер.с англ. Д. І. Арнольда. – М.: Мир, 1978. – 308 с. 5. Отт, Г. Методи придушення шумів і перешкод в електронних системах / Г. Отт; пер. з англ. Б. Н. Броніна; під ред. М. В. Гальперіна. – М.: Мир, 1979. – 318 с.

2. Михайлов, Є. В. Перешкодозахищеність інформаційно-вимірювальних систем / Є. В. Михайлов. – М.: Енергія, 1975. – 312 с.

3. Шахов, Е. К. Підвищення завадостійкості цифрових засобів вимірювання / Е. К. Шахов. – Пенза: ППІ, 1983. – 48 с.

4. Методи автоматичного розпізнавання мови: в 2 кн. : Пров. з англ. / Д. Х. Клетт, Дж. А. Барнет, М. І. Бернстейн і ін.; під ред. У. Лі. – М.: Мир, 1983. – Кн. 2. – 392 с.

5. Методи автоматичного розпізнавання мови: в 2 кн. : Пров. з англ. / У. А. Лі, Е. П. Нейбург, Т. Б. Мартін і ін.; під ред. У. Лі. – М.: Мир. 1983. – Кн. 1. – 328 с.

6. Дігун, О. Г. Сигнали, перешкоди, шуми: навч. посібник / О. Г. Дігун, В. І. Веприк. – Новочеркаськ: НГТУ, 1994. – 94 с.

7. Болл, Р. М. Керівництво по біометрії / Р. М. Болл, Дж. Х. Коннел, Н. К. Ратха; пер з англ. Н. Е. Агапова. – М.: Техносфера, 2007. – 352 с. 13. Фролов, А. В. Синтез і розпізнавання мови. Сучасні рішення / Г. В. Фролов. – М.: Зв'язок, 2003. – 216 с.

8. Алейник С. В., Матвеев Ю. Н., Раев А. Н. Метод оценки уровня клиппирования речевого сигнала // Науч.техн. вестн. информационных технологий, механики и оптики. 2012. № 3 (79). С. 79 – 83.

9. Белых И. Н., Капустин А. В., Козлов А. В., Лоханова А. И., Матвеев Ю. Н., Пеховский Т. С., Симончик К. К., Шулипа А. К. Система идентификации дикторов по голосу для конкурса NIST SRE 2010 // Информатика и ее применения. 2012. Т. 6, № 1. С. 91 – 98.

10. Рабинер, Л. Р. Цифровая обработка речевых сигналов / Л. Р. Рабинер, Р. В. Шафер. – М. : Радио и связь, 1981. – 496 с.

11. S. Cho, C. Han, D. H. Han, and H.-I. Kim. Web-based keystroke dynamics identity verification using neural network // Journal of Organizational Computing and Electronic Commerce. 2000. №10 (4). P. 295-307.

12. D.-T. Lin. Computer-access authentication with neural network based keystroke identity verification // International Conference Neural Networks. 1997. №1. P. 174-178.

13. M. Obaidat, Sadoun B. Verification of computer users using keystroke dynamics // in Systems, Man, and Cybernetics, Part B: Cybernetics. 1997. №27(2). P. 261-269.

14. F. Monrose, Rubin A.D. Authentication via keystroke dynamics // in Proceedings of the 4th ACM conference on Computer and communications security. New York, NY, USA: ACM, 1997. №13. P. 48-56.

15. M. Obaidat, Macchiarolo D. An online neural network system for computer access security // in Industrial Electronics, IEEE Transactions. 1993. № 40(2). P. 235-242. Kawalerowicz, M. and Berntson, C. Continuous Integration in .NET. — Manning, 2011. — 303 p.

16. Nechiporenko A.S., Gubarenko E.V., Gubarenko M.S. (2019) Authentication of users of mobile devices by their motor reactions. Telecommunications and Radio Engineering, 78 (11). pp. 987-1003. DOI: 10.1615/TelecomRadEng.v78.i11.60

17. PRIORITY DIRECTIONS OF SCIENCE AND TECHNOLOGY DEVELOPMENT Abstracts of I International Scientific and Practical Conference Kyiv, Ukraine 27-29 September 2020, p. 207.

18. Galunov V.I. 2007. Verifikacija i identifikacija govorjashhego. S-Peterburgskij gosudarstvennyj universitet, p.201.

19. Мясников Л.Л., Мясникова.Н.Е. 1970. Автоматическое распознавание звуковых образов Л., «Энергия», 183. Mjasnikov L.L., Mjasnikova N.E. 1970. Avtomaticheskoe raspoznavanie zvukovyh obrazov L., «Jenergija», p. 183.

20. Вологдин Э.И. 2004. Слух и восприятие звука: Учеб. пособие. СТ«Факультет ДВО» , СПб. Vologdin Je.I. 2004. Sluh i vosprijatie zvuka: Ucheb. posobie. ST«Fakul'tet DVO» , p.132.

21. Chernomorec A.A. Prohorenko E.I., Goloshhapova A.A., 2009. About properties of subband matrices eigenvectors. Nauchnye vedomosti BelGU. Istoriya. Politologiya. Ekonomika. Informatika. [Belgorod State University Scientific Bulletin. History Political science Economics Information technologies], p. 122-128.

22. Selina Chu and Eamonn Keogh and David Hart and Michael Pazzani Iterative Deepening Dynamic Time Warping for Time Series, In Proc 2 nd SIAM International Conference on Data Mining, 2002.

23. Ann Chotirat and Ratanamahatana Eamonn and Keogh Everything you know about Dynamic Time Warping is Wrong, The 31st Annual International Symposium on Forecasting, 2004.

24. E.G. Caiani and A. Porta and G. Turiel and M. Muzzupappa and S. Pieruzzi and F. Grema and C. Malliani and A. Cerutti and S. Cerutti Warped-average template technique to track on a cycleby-cycle basis the cardiac filling phases on left ventricular volume, IEEE Computers in Cardiology, 1998.

25. Donald J. Berndt and James Clifford Using Dynamic Time Warping to Find Patterns in Time Series, KDD Workshop, 1994.

26. Georgios N. Banavas and Sue Denham and Michael J. Denham Fast Nonlinear Deterministic Forecasting Of Segmented Stock Indices Using Pattern Matching And Embedding Techniques, Society for Computational Economics, 2000.

27. 27. John Aach and George M. Church Aligning Gene Expression Time Series With Time Warping Algorithms, 2001.