МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний аерокосмічний університет ім. М.Є. Жуковського
«Харківський авіаційний інститут»

Факультет програмної інженерії та бізнесу

Кафедра інженерії програмного забезпечення

Пояснювальна записка до дипломної роботи

магістра
(освітній ступінь)

на тему «Експериментальне дослідження ефективності методів розпізнавання мови»

ХАІ.603.667п1.121.156333.20О

Виконав: студент 6 курсу групи № 667-п1
Спеціальність 121 – Інженерія програмного забезпечення
(код та найменування)
Освітня програма Хмарні обчислення та Інтернет речей
(найменування)
Гур`єв В.А.
(прізвище й ініціали студента)
Керівник: Мандрікова Л.В.
(прізвище й ініціали)
Рецензент: Ільїна І.В.
(прізвище й ініціали)

Харків – 2020

Міністерство світи і науки України
Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»

Факультет програмної інженерії та бізнесу
(повне найменування)
Кафедра інженерії програмного забезпечення
(повне найменування)
Рівень вищої освіти другий (магістерський)
Спеціальність 121 – інженерія програмного забезпечення
(код та найменування)
Освітня програма хмарні обчислення та Інтернет речей
(найменування)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ І. Б. Туркін
(підпис) (ініціали та прізвище)
“_____” _____ 2020 року

З А В Д А Н Н Я
НА ДИПЛОМНУ РОБОТУ СТУДЕНТУ

Гур’єву Вадиму Андрійовичу
(прізвище, ім’я, по батькові)

1. Тема дипломної роботи Експериментальне дослідження ефективності методів розпізнавання мови

керівник дипломної роботи Мандрікова Людмила Василівна, доц. каф. 603
(прізвище, ім’я, по батькові, науковий ступінь, вчене звання)
затверджені наказом Університету № _____ від “____” _____ 2020 року
2. Термін подання студентом роботи _____
3. Вихідні дані до роботи Експериментальне дослідження ефективності методів розпізнавання мови

4. Зміст пояснювальної записки (перелік питань, які потрібно розробити) ____
Постановка задачі та аналіз предметної області з розпізнавання голосових сигналів
Планування експериментальних досліджень ефективності методів розпізнавання мови
Аналіз результатів експериментального дослідженя та формування практичних рекомендацій щодо ефективності методів розпізнавання мови

5. Перелік графічного матеріалу Пояснювальна записка – 1 (усього – 77 сторінки), рисунків – 17, таблиць – 15.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1	Мандрікова Л.В., доцент каф. 603		
2	Мандрікова Л.В., доцент каф. 603		
3	Мандрікова Л.В., доцент каф. 603		

Нормоконтроль _____ ____В. А. Постернакова_____ «___»_____2020р.
(підпис) (ініціали та прізвище)

7. Дата видачі завдання «_9_» __вересня__ 2020 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломної роботи	Строк виконання етапів роботи	Примітка
1	Отримання і затвердження теми дипломного проекту	04.09.2019 – 15.09.2019	
2	Аналіз предметної області	16.09.2019 – 25.09.2019	
3	Постановка задачі	26.09.2019 – 30.10.2019	
4	Проведення теоретичних досліджень	01.11.2019 – 23.11.2019	
5	Планування експериментальних досліджень	13.01.2020 – 18.05.2020	
6	Обробка та аналіз експериментальних досліджень, формування практичних рекомендацій	19.05.2020 – 05.09.2020	
7	Підготовка пояснювальної записки	06.09.2020 – 29.09.2020	
8	Оформлення пояснювальної записки до дипломного проекту	01.10.2020 – 26.11.2020	
9	Предзахист дипломного проекту	27.11.2020	
10	Захист дипломного проекту	03.12.2020	

Студент _____ ____Гур'єв В. А.__
(підпис) (прізвище та ініціали)

Керівник проекту ____ _Мандрікова Л.В._
(підпис) (прізвище та ініціали)

РЕФЕРАТ

Дипломний робота: 77 с., 17 рис., 15 табл., 29 джерел.

Метою дипломної роботи магістра є експериментальне дослідження та формування рекомендацій щодо підвищення ефективності методів розпізнавання голосових сигналів. Я запропонував часткове рішення цієї проблеми: модифікувати модель алгоритму динамічної трансформації часової шкали голосових сигналів за рахунок зменшення еталонів, для подальшого розвитку методу динамічної трансформації часової шкали в системах розпізнавання голосових сигналів за рахунок використання запропонованої комбінованої моделі. Для досягнення мети необхідно вирішити наступні завдання:

- проаналізувати сферу «Методи розпізнавання мови»;
- проаналізувати та відібрати методи розпізнавання мови;
- проаналізувати існуючі програм для розпізнавання мови;
- спланувати експериментальне дослідження вибраних методів;
- проаналізувати результатів проведеного експерименту;
- сформувати практичні рекомендації для удосконалення методів розпізнавання мови завдяки запропонованої комбінованої моделі.

Актуальність завдання зумовлена все більшою популярністю систем розпізнавання мовлення, отже удосконалений метод покращить ефективність існуючих алгоритмів. Метою також є формування практичних рекомендацій щодо створення комбінованої моделі.

Основний практичний результат роботи полягає в тому, що запропонована модифікація алгоритму динамічної трансформації часової шкали та комбінована модель, наведена порівняльна характеристика існуючих методів розпізнавання мови, а також отримані результати можуть бути використовуватись у майбутніх дослідженнях за даним напрямком.

МОВА, ГОЛОС, АЛГОРИТМ, НЕЙРОМЕРЕЖА, РОЗПІЗНАВАННЯ ГОЛОСОВИХ СИГНАЛІВ

РЕФЕРАТ

Дипломный проект: 77 с., 17 рис., 15 табл., 29 источника.

Целью дипломной работы магистра является экспериментальное исследование и формирование рекомендаций по повышению эффективности методов распознавания голосовых сигналов. Я предложил частичное решение этой проблемы: модифицировать модель алгоритма динамической трансформации временной шкалы голосовых сигналов за счет уменьшения эталонов, для дальнейшего развития метода динамической трансформации временной шкалы в системах распознавания голосовых сигналов за счет использования предложенной комбинированной модели. Для достижения цели необходимо решить следующие задачи:

- проанализировать сферу «Методы распознавания речи»;
- проанализировать и отобрать методы распознавания речи;
- проанализировать существующие программ для распознавания речи;
- спланировать экспериментальное исследование выбранных методов;
- проанализировать результатов проведенного эксперимента;
- сформировать практические рекомендации по совершенствованию методов распознавания речи благодаря предложенной комбинированной модели.

Актуальность задачи обусловлена все большей популярностью систем распознавания речи, следовательно, усовершенствованный метод улучшит эффективность существующих алгоритмов. Целью также является формирование практических рекомендаций по созданию комбинированной модели.

Основной практический результат работы состоит в том, что предложенная модификация алгоритма динамической трансформации временной шкалы и комбинированная модель, приведенная сравнительная характеристика существующих методов распознавания речи, а также полученные результаты могут быть использованы в будущих исследованиях по данному направлению.

ЯЗЫК, ГОЛОС, АЛГОРИТМ, НЕЙРОСЕТИ, РАСПОЗНАВАНИЕ ГОЛОСОВЫХ СИГНАЛОВ

ABSTRACT

Master’s thesis: 77 pages, 17 figures, 15 tables, 29 references.

A goal of the master's diploma is an experimental study and the formation of recommendations for improving the efficiency of voice signal recognition methods. I proposed a partial solution to this problem: to modify the model of the algorithm of dynamic transformation of the time scale of voice signals by reducing the standards, to further develop the method of dynamic transformation of the time scale in voice recognition systems using the proposed combined model.

For reaching of the goal it is necessary to solve the following problem:

- to analyze the scope of «Language Recognition Methods»;
- to analyze and select methods of language recognition;
- to analyze existing language recognition programs;
- to plan an experimental study of selected methods;
- to analyze the results of the experiment;
- to form practical recommendations for improving language recognition methods due to the proposed combined model.

The urgency of the task is due to the growing popularity of speech recognition systems, so the improved method will improve the efficiency of existing algorithms.

A key objective is also to form practical recommendations for creating a combined model.

The main practical result of the work is that the proposed modification of the algorithm of dynamic transformation of the time scale and the combined model, a comparative description of existing methods of speech recognition, and the results can be used in future research in this area.

LANGUAGE, VOICE, ALGORITHM, NEURAL NETWORK, VOICE SIGNAL RECOGNITION

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

Алгоритм динамічної трансформації часової шкали (англ. dynamic time warping, DTW) є одним з алгоритмів вимірювання подібності між двома тимчасовими послідовностями, які можуть різнитися за швидкістю.

Апроксимація − науковий метод, що полягає в заміні одних об'єктів іншими, в якомусь сенсі близькими до вихідних, але більш простими.

АРМ − автоматичне розпізнавання мовлення − процес перетворення мовленнєвого сигналу в текстовий потік.

АЦП − аналого-цифровий перетворювач − пристрій, що перетворює вхідний аналоговий сигнал в дискретний код (цифровий сигнал), який кількісно характеризує амплітуду вхідного сигналу.

АЧХ − амплітудно-частотна характеристика − залежність амплітуди вихідного сигналу пристрою або системи передачі, підсилення або обробки сигналу від частоти вхідного сигналу сталої амплітуди.

ВСШ − Відношення сигнал-шум (англ. Signal-to-noise ratio, SNR) − безрозмірна величина, що дорівнює відношенню потужності корисного сигналу до потужності шуму.

ВСТУП

Важливим завданням розробки технічних систем є забезпечення інтуїтивного і природного інтерфейсу з користувачем, оскільки сучасні комп'ютерні програми орієнтовані на користувачів і розвиваються відповідно до їх зростаючих потреб.

Однією з природних форм взаємодії для людини є мова. Голосовий інтерфейс користувача спроможний забезпечити зручний і гнучкий спосіб взаємодії людини з комп'ютером, оскільки для його використання не потрібно опановувати новими навичками.

Голосовий інтерфейс якісним чином змінює спосіб, а отже і ефективність взаємодії користувача з системою. Голосовий пошук від компанії Google і голосовий асистент Siri від компанії Apple є цьому яскравими прикладами, підтверджуючи нагальну необхідність впровадження мовних технологій, зокрема розпізнавання і синтез мови.

Складність розпізнавання мови полягає в тому, що сукупність таких характеристик голосу і мови як тембр, гучність, висота, темп, інтонація, якість дикції роблять мову кожної людини неповторною і унікальною як відбитки пальців. Завданням комп'ютерної техніки та програмного забезпечення є розпізнавання сказані людиною слова в будь-яких умовах без попередньої адаптації під конкретний голос.

Актуальність роботи. Зараз систем розпізнавання мовлення набувають все більшої популярності та зустрічаються все частіше. Успішними прикладами використання технології розпізнавання мови в мобільних додатках є: введення адреси голосом в Яндекс - Навігатор, голосовий пошук Google Now.

Крім мобільних пристроїв, технологія розпізнавання мови знаходить широке поширення в різноманітних сферах людської діяльності. Телефонія – автоматизація обробки вхідних і вихідних дзвінків шляхом створення голосових систем самообслуговування зокрема для: отримання довідкової інформації та консультування, замовлення послуг, товарів, проведення опитувань, анкетування,

збору інформації, голосове керування побутовою технікою, голосове управління в салоні автомобіля - наприклад, навігаційною системою, соціальні сервіси для людей з обмеженими можливостями, комплексні системи захисту інформації. Голосова аутентифікація, визначення емоційного забарвлення голосу диктора.

Задачі дослідження:

- проаналізувати сферу «Методи розпізнавання мови»;
- проаналізувати та відібрати методи розпізнавання мови;
- проаналізувати існуючі програм для розпізнавання мови;
- спланувати експериментальне дослідження вибраних методів;
- проаналізувати результатів проведеного експерименту;
- сформувати практичні рекомендації для удосконалення методів розпізнавання мови завдяки запропонованої комбінованої моделі.

Об’єктом дослідження є процес розпізнавання голосових сигналів.

Предметом дослідження є методи та моделі розпізнавання голосових сигналів.

Мета роботи: експериментальне дослідження та формування рекомендацій щодо підвищення ефективності методів розпізнавання голосових сигналів.

Методи дослідження. В роботі використовуються методи планування експерименту, статистичний аналіз результатів експерименту, математичного моделювання, методи оптимізації, методи системного аналізу.

Наукова новизна роботи полягає в модифікації моделі алгоритму динамічної трансформації часової шкали голосових сигналів за рахунок зменшення еталонів, для подальшого розвитку методу динамічної трансформації часової шкали в системах розпізнавання голосових сигналів за рахунок використання запропонованої комбінованої моделі.

Практична цінність отриманих в роботі результатів полягає в тому, що запропонована модифікація алгоритму динамічної трансформації часової шкали наведена порівняльна характеристика існуючих методів розпізнавання мови, а також отримані результати можуть бути використовуватись у майбутніх дослідженнях за даним напрямком, враховуючи переваги та недоліки.

1 ПОСТАНОВКА ЗАДАЧІ ТА АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ З РОЗПІЗНАВАННЯ ГОЛОСОВИХ СИГНАЛІВ

Постановка задачі

Задача розпізнавання мови полягає в точному і ефективному, в контексті алгоритму класифікації, відтворенні вимовленого мовленнєвого сигналу. В підходах, що використовуються сьогодні, її рішення полягає в послідовному порівнянні з еталонами, що задані словником системи розпізнавання мови. Звісно словником можуть виступати різноманітні фонеми природної мови, що робить можливим побудову системи розпізнавання мови навіть без словника у прямому розумінні цього слова. Словник може лише допомагати виправити помилки розпізнавання.

З точки зору інформаційної ентропії, задача побудови ефективної дикторонезалежної стратегії розпізнавання мови може бути сформульована як задача пошуку оптимального за загальносистемним критерієм дерева рішень, в якому на кожному кроці класифікації з апріорного алфавіту вибирають підмножину ознак, що максимально зменшує на досягнутому кроці ентропію про образ і збільшує швидкість класифікації. Така стратегія передбачає використання множинного описання слів у термінах різних фонетичних класів, що відповідають різним рівням дерева класифікації, а також вибору інформативних дикторонезалежних ознак для виділення фонетичних класів на кожному рівні.

Коли говорять про розпізнавання мовленнєвої інформації, об’єктом дослідження вважають власне процес оброблення інформації про мовленнєві сигнали, а предметом дослідження вважається технологія аналізування, оброблення і параметризації мовленнєвої інформації. Процеси видалення шуму, сегментації і виявлення вокалізованих ділянок є власне обробкою мовленнєвих сигналів, тобто також є предметом дослідження.

Класифікація систем розпізнавання мови

Автоматизовані системи розпізнавання мови можуть бути класифіковані

за багатьма ознаками: за типом мови, за множиною дикторів, за об’ємом та за повнотою словника, що необхідно розпізнавати.

За типами мову поділяють на дискретну і неперервну. Дискретною називають таку мову в якій паузи між словами значно більші за природні паузи всередині слів, наприклад, надиктування окремих команд. Неперервна мова в свою чергу не має значної паузи між словами. Природний людський режим спілкування є неперервною мовою.

За ознакою множини дикторів системи розпізнавання мови поділяють на дикторозалежні, тобто такі, якість розпізнавання яких залежать від індивідуальних особливостей вимовляння диктора та дикторонезалежні. Дикторозалежні системи розпізнавання звичайно розпізнають мовленнєву інформацію сказану не тільки одним конкретним диктором, просто ймовірність вдалого розпізнавання слова, сказаного одним конкретним диктором, вища за середню ймовірність вдалого розпізнавання цього слова, сказаного іншими дикторами. У відповідності до цього положення дикторонезалежною називають таку систему, ймовірність вдалого розпізнавання слова якою однакова для усіх дикторів.

За об’ємом словника системи розпізнавання мови класифікуються на дві категорії: системи з малими та системи з великими словниками. Ці системи значно відрізняються одна від одної. Так, систему з малим словником можна навчити послідовно, вимовляючи кожне слово зі словнику.

Систему з великим словником необхідно навчати синтезованими акустичними ознаками слів (фонеми, трифони), адже в цьому випадку неможливо надиктувати системі весь словник.

Повнота словника полягає в тому, що кожне слово вхідного мовленнєвого сигналу має бути присутнім в словнику. Як правило повний словник мають лише системи розпізнавання мови з малим словником.

Первинна обробка мовних сигналів

Первинна обробка мовних сигналів полягає у видаленні шумів, виокремлення значимої для розпізнавання інформації, усунення варіантності

диктора і навколишнього середовища, стиснення сигналу, сегментація на фонеми.

Методи аналізу мовних сигналів, тобто безпосереднього виділення ознак, поділяються на три групи [1]:

1) Фонетичні методи, які спираються на теорію мовотворення. Суть їх полягає у виділенні ознак, що характеризують спосіб артикуляції. Фонетичні методи аналізу мовних сигналів можуть розглядатися як перший рівень розпізнавання мови, тому що більшість з них засновано на деяких перетвореннях первинних ознак мовних сигналів.

2) Неакустичні методи, що по своїй меті примикають до фонетичних і полягають у виокремленні інформації про процеси, що супроводять артикуляцію.

3) Параметричні методи, що засновані, по-перше на представленні мовного сигналу як реалізації деякого процесу в часі і, по-друге, виокремлення деяких параметрів цього процесу, найчастіше пов'язаних з його спектральними характеристиками. Відомими параметричними методами аналізу мовленнєвої інформації є:

- спектрально-смугові, кореляційні і ортогональні методи;

- цифрові фільтри; — методи обчислення спектра за допомогою швидкого перетворення Фур'є;

- методи, що застосовують вейвлет-перетворення для моделювання мовних сигналів;

- методи лінійного передбачення мови (ЛПМ);

- методи, пов'язані з виділенням миттєвої частоти переходів через нуль;

- часові методи, засновані на аналізі розподілу тривалості інтервалів між переходами через нуль або екстремумами мовного сигналу;

- використання нелінійного перетворення і фазові співвідношення мовного сигналу.

Більшість з цих методів заснована на загладжувані сигналу та обчисленні спектру, чи кепстру.

Як правило, якісна первинна обробка мовленнєвої інформації включає в себе 5 етапів: видалення шуму, сегментація, виокремлення вокалізованих ділянок, вимір частоти основного тону та параметризація.

Існує багато методів, що реалізують ці етапи. Деякі методи можуть одночасно реалізовувати декілька етапів первинної обробки сигналів.

Для розпізнавання мови, як ізольованої, так і злитої, необхідно заздалегідь визначити її межі в контексті навколишніх даремних звукових сигналів. Складність визначення меж мови пов'язана з особливостями вимови конкретного диктора, наявністю в мовному сигналі різних видів сторонніх шумів, а також звукових артефактів процесу артикуляції (придих, чмокання і т.п.).

Методи, засновані на обчисленні короткочасної енергії сигналу, спектральної енергії та кількості нуль-перетинів нестійко працюють в умовах, коли з'являються шуми з динамічним спектром або відносно сильні стаціонарні шуми.

В будь-якій мові існує деякий набір звуків, який бере участь при формуванні звукового обрису слів. Як правило, звук поза мовою не має значення, він набуває його лише як складова частина слова, допомагаючи відрізнити одне слово від іншого. Елементи цього набору звуків називаються фонемами.

Звуки, що беруть участь у формуванні мови, мають дві основні класифікації: за ознаками артикуляцій і за акустичними ознаками.

Класифікація звуків за ознаками артикуляцій (екскурсія, витримка, рекурсія) є вкрай важливою при використанні методів розпізнавання мови за допомогою моделювання носоглотки, але для вирішення завдань ділення на фонеми цікавіший розгляд акустичних відмінностей звуків. За акустичними ознаками звуки підрозділяються на:

Тональні звуки — утворюються голосом при повній відсутності шумів, що забезпечує хорошу чутність звуку. Тональними звуками є усі голосні.

Сонорні (звучні) — чия якість визначається характером звучання голосу,

який грає головну роль в їх формуванні, а шум бере участь в мінімальному ступені. Сонорними приголосними є: й, л, р.

Шумні — їх якість визначається характером шуму. Шумні приголосні поділяються на дзвінкі тривалі (в, з, ж), дзвінкі миттєві (вибухові) (б, д, г), глухі тривалі (ф, с, ш, х) та глухі миттєві (вибухові) (п, т, к).

Різниця між звуками різних видів є дуже великою і теоретично, це значно полегшило-б завдання поділу звуків, але в реальних умовах, коли ми маємо справу зі злитою мовою (мовним ланцюгом), етап екскурсії артикуляції наступного звуку накладається на рекурсію, чи навіть на етап витримки артикуляції попереднього звуку, що нівелює цю різницю.

Видалення шуму з мовних сигналів

В загальному випадку видалення шуму полягає в виокремленні значущого мовного сигналу від фонового шуму, та видаленні визначеного шуму з вхідного сигналу. Фоновий шум буває стаціонарний і нестаціонарний.

В [2] ділянки мовленнєвого сигналу, на які накладено стаціонарний шум знаходять за допомогою дискретного перетворення Фур’є та імовірнісних розподілів шуму та мови. Спочатку записується тільки шум і оцінюються параметри його рівномірного та Гауссівського розподілу. Потім, за допомогою порогової умови на щільність ймовірності нормального розподілу, із вхідного мовного сигналу виділяють ділянки з можливою сумішшю корисного сигналу і шуму. Для отриманих ділянок сигналу знову оцінюють параметри Гауссівського розподілу і, ґрунтуючись на теоремі Байеса, приймають рішення про наявність корисного сигналу в виокремленній ділянці.

Виокремлення ділянок з корисним сигналом, на який накладається нестаціонарний шум в [2] вирішують представленням шуму і мови в якості статистичних моделей прихованої Марківської мережі.

У загальному випадку фільтрація сигналу полягає в тому, щоб виокремити корисну складову сигналу і видалити сторонні шуми і спотворення. В залежності від характеру шуму виникає декілька завдань:

- необхідно виокремити корисний сигнал з високочастотного або

смугового шуму. В цьому випадку фільтрація полягає у виборі типу фільтру і розрахунку його параметрів [3];

- потрібно виокремити мовний сигнал з мово-подібного шуму. Наприклад, два або більше дикторів можуть говорити одночасно, а потрібно отримати розбірливу мову тільки одного з них. Це одне з найскладніших завдань фільтрації, загальних методів рішення якої поки не існує;

- потрібно відновити сигнал, що зазнав нелінійних спотвореннь. Це завдання виникло з появою цифрових телефонних ліній, які ущільнюють і тим самим спотворюють початковий сигнал.

Видалення стаціонарного шуму в [2] реалізують фільтрацією компонент ДПФ вхідного сигналу синтезованими нерекурсивними фільтрами.

Задача видалення нестаціонарного шуму, в загальному випадку, поділяється на два великих класи задач, в залежності від того, чи відома нам заздалегідь фізична модель нестаціонарного шуму.

Перший напрям ґрунтується на тому, що інформації про фізичний процес генерації нестаціонарного шуму досить для побудови його моделі, що дозволяє за деякими параметрами відрізняти компоненти шуму від компонентів корисного сигналу.

Другий напрям використовують в випадках, коли фізичну модель шуму побудувати не можливо. Розвиток цього напряму ґрунтується на тому, що фільтр проходить стадію навчання, де за допомогою реалізацій незашумленої мови створюються її стани. Процес фільтрації схожий з процесом розпізнавання, де визначається стан мови, з максимальною ймовірністю схожий на ділянку зашумленої мови, і який згодом замінює собою цю ділянку. Оскільки варіативність мови від диктора до диктора висока, то така фільтрація припускає або шумоочистку сигналу наперед відомого диктора, яким і проводилося навчання фільтру, або створення достатньо великого банку голосів дикторів і відповідних ним станів в надії, що вдасться перебрати всі типи голосів.

В [2] можна знайти моделі для видалення деяких нестаціонарних шумів

першого класу

Сегментація мовних сигналів

Сегментація мовного сигналу полягає у виділенні ділянок сигналу, що відповідають окремим структурним одиницям мовного сигналу. Якщо в якості таких одиниць розглядати фонеми, то завдання сегментації зводиться до виявлення міжфонемних переходів. В рамках традиційних підходів розв'язання цієї задачі вельми проблематично.

В системах розпізнавання мови для визначення меж мови традиційно використовуються методи (наприклад, Voice Activity Detector), засновані на обчисленні короткочасної енергії сигналу або спектральної енергії. Крім того, додатково застосовуються методи, що використовують кількість нуль-перетинів сигналу і інформацію про тривалість мовних фрагментів. Недоліком цих алгоритмів є ненадійності в умовах нестаціонарного шуму, а також при виникненні різних звукових артефактів (придих, чмокання і таке інше). Також існують алгоритми, засновані на адаптивних порогових значеннях, але при виникненні звукових артефактів, а також відносно високому рівні шуму або незначному рівні корисного сигналу вони також стають нестійкими.

Надійний метод сегментації мовленнєвої інформації має задовольняти такі вимоги:

- забезпечення мінімальної вірогідності помилкового спрацювання при дії тільки шуму з високим рівнем;

- висока вірогідність правильного виділення мови навіть в умовах сильного шуму;

- висока швидкодія для виключення затримок включення і виключення алгоритму розпізнавання мови.

Для практичних цілей кожна фонема може бути представлена квазістатичним спектром, в якому передавальна функція, не змінюється в часі [4]. Явища, що обумовлені швидкими змінами функції джерела, можуть служити для розмежування окремих фонем в мовному потоці. Різкі зміни передівальної функції, що пов'язані з швидкою зміною положення

артикулюючих органів, також вказують на межу (початок або кінець) фонеми. Звичайно, мінімальна швидкість зміни повинна бути визначена експериментально для кожного випадку. Додатковим засобом для визначення межі фонеми є швидкі флуктуації загальної інтенсивності звукової хвилі.

1.1 Загальний опис проблеми розпізнавання голосових сигналів

Проблема створення усного діалогу людини з машинами є однією з найбільш актуальних проблем кібернетики, інформатики і обчислювальної техніки. Оснащення ЕОМ засобами розпізнавання та синтезу мови має та ще в більшій степені буде мати велике економічне та соціальне значення. Це забезпечить доступність ЕОМ всьому населенню, можливість програмування і рішення задач на природній мові, безпаперову технологію управління, скорочення термінів навчання користувачів ЕОМ, підвищення продуктивності праці в сферах виробництва, розподілення і в побуті, підвищення ефективності використання техніки, створення сприятливих умов праці.

Таким чином, мова йде про створення й використання інтелектуальних ЕОМ і людино-машинного інтерфейсу на природній мові в комп’ютерних системах.

Історія науки і техники налічує чимало спроб створення «слухаючих» та «говорящих» машин починаючи ще з XVIII століття. Цьому в значній степені сприяли становлення та розвиток електроніки та електрозв’язку. Але найбільший інтерес до проблеми та її розвиток починаються одночасно з появою ЕОМ і їх широким розповсюдженням, з автоматизацією різних областей діяльності людини.

Усний діалог людини з ЕОМ в найбільш зручній та звичній для людини формі – голосом – став техніко-економічною і соціальною необхідністю. Чисто в науковому плані кінцевою метою досліджень є створення засобів усного діалогу людини і ЕОМ на природніх мовах, наприклад автоматичною машинки, шо друкує та редагує тексти під диктовку, або машин-перекладачів

з голосу.

Процес розпізнавання мови являє собою перетворення акустичного сигналу, отриманого від мікрофона, в послідовність слів. Отриманий набір гіпотез ланцюжків слів далі використовується для розуміння мови. При цьому виникає ряд проблем. По-перше, людина зазвичай не робить паузи між словами, а при злитому проголошенні до задачі розпізнавання додається ще й завдання виділення слів з потоку мови, що свідомо більш складно. Виникає необхідність виділяти односкладові слова - саме з ними і пов'язано максимальне число помилок реально існуючих систем. Можна вимагати, щоб людина вимовляв слова по одному, роблячи досить тривалі паузи або щоб кожне наступне слово вимовлялося після звукового сигналу. Але даний підхід не зручний і може бути застосований лише для подачі простих команд.

Наступна проблема - різниця голосів, діалектів, дикція, вікових відмінностей, емоційний і фізичний стан диктора. Значний вплив вносить акустичний аспект, тобто зміна мікрофона, розташування мікрофона щодо рота, акустична обстановка в приміщенні.

Саме через ці та багатьох інших проблем до повного вирішення задачі розпізнавання мови і раніше досить далеко. Існує два істотно розрізняються режими роботи: з налаштуванням на голос певного диктора і без такого налаштування. Розміри словника при роботі з налаштуванням на диктора (speaker-dependent) в даний час можуть досягати декількох (і навіть багатьох) тисяч слів при злитому проголошенні. Процедура настройки на диктора виглядає наступним чином: диктор читає якийсь спеціальним чином складений текст, комп'ютер розпізнає слова і видає варіант розпізнавання. Диктор позначає помилки і читає текст знову. Після кількох таких ітерацій процес сходиться, і комп'ютер виявляється в стані розпізнавати мовлення.

Нарешті, останній, найбільш складний для реалізації, але водночас і найбільш перспективний режим роботи - розпізнавання без настройки на диктора. При цьому гарантується, що система розпізнає будь-яке включене в словник слово, ким би воно не було вимовлено. Тут, як правило, словники

налічують невелику кількість слів (зазвичай не більше двох десятків) і існують для відносно невеликого числа мов (приблизно тридцяти). Українська мова в це число хоча і входить, проте кількість розпізнаваних українських слів невелика.

Створення словника для розпізнавання мови без настройки на голос вимагає великих витрат. Для вирішення цього завдання розробникам доводиться опитувати велику кількість (кілька сотень або тисяч) носіїв мови, виділяти якісь загальні елементи мови, усереднювати їх - і все це для того, щоб забезпечити розпізнавання десяти-двадцяти слів. Найчастіше словник без настройки на голос користувача вимагає роздільного проголошення слів. Для цілого ряду додатків цього, однак, виявляється цілком достатньо.

Розпізнавання мовлення часто називають терміном «розпізнавання мови». Це не зовсім правильно, оскільки існує окрема задача розпізнавання мови, що передбачає відповідь на запитання, якою мовою розмовляє користувач, якого ми називатимемо терміном «диктор». Інколи вживається термін «розпізнавання голосу». Це може означати і введення тексту голосом, і ідентифікацію людини за голосом, і виділення голосових сегментів у звуковому сигналі.

Загалом, метою розпізнавання мовлення є отримання різного роду інформації на основі вхідного мовленнєвого (голосового) сигналу: про що говориться, хто говорить, якою мовою, в якому фізичному стані перебуває диктор тощо.

Ось доволі вичерпний перелік проблем, які вирішуються в ділянці розпізнавання та розуміння мовлення:

- автоматичне перетворення мовленнєвого сигналу на текст;
- введення інформації голосом, диктувальна машина;
- пошук ключових слів і фраз у потоці мовлення;
- смислова інтерпретація голосових повідомлень;
- ідентифікація та верифікація диктора;
- адаптація до голосу диктора та акустичного каналу;

- розпізнавання мови, якою говорить диктор, його акценту;

- усний переклад з однієї мови на іншу;

- розпізнавання емоційного та фізичного стану мовця.

Завдяки розпізнаванню мовлення вивільняються руки користувача при керуванні комп’ютерними системами, введенні текстової інформації, транскрибуванні (стенографуванні) фонограм тощо. Вже тепер починають з’являтися системи, що допомагають в оволодіванні розмовною іноземною мовою на основі технології розпізнавання мовлення.

Якщо поруч із звуковим нам доступний зоровий канал, то його можна використовувати як додаткову інформацію при вирішенні наведених задач. В такому разі йдеться про технології мультимодального розпізнавання та розуміння мовлення. А при поєднанні технологій розуміння мовлення та синтезу мовлення за текстом виникає система усного діалогу

1.2 Аналіз задачі виділення параметрів звукового сигналу та подальшого розбору

Розпізнавання голосу − процес трансформації мовного сигналу в цифрові повідомлення, наприклад текстову інформацію. Мовний сигнал − це звук, синтезований людським апаратом мови [2].

Як відомо, звук в комп'ютерній техніці може відображатися в вигляді деякого набору його амплітуд, вироблених через деякі проміжки часу (період дискретизації) і подаються деякою кількістю розрядів у двійковій системі числення (розрядність вибірки). Такий формат зберігання звукового сигналу зручний для його перетворення назад в статичний сигнал. Однак, деякі маніпуляції зі звуковим сигналом збереженим в такому форматі буває не зручно, або взагалі немає можливості здійснювати. Це пояснюється тим, що у реальності звуковий сигнал складається із певних його частот з певною фазою і амплітудою. Таким чином, застосовуючи такі методи розпізнавання звукового сигналу як "фільтрування нижніх частот", "фільтрування верхніх

частот", або процес обробки голосових сигналів, треба трансформувати формат звукового сигналу у вигляд відліків його частотного спектра. Після цієї трансформації звуковий сигнал буде представлений у чисельному вигляді, відповідно до амплітудних і фазово-частотних складових [3].

Основними цілями при розпізнаванні мовних сигналів є:

- виділення параметрів вхідного сигналу;
- виділення придатних для розбору характеристик із вхідного звукового сигналу;
- розбір сигналу з виділеними параметрами − пошук закономірності між вхідним голосовим сигналом та вимовленим словом.

Ці задачі являються основними як для існуючих, застарілих систем, так і для нових систем.

1.2.1 Задача трансформації вхідного мовного сигналу у формат придатний для подальшого аналізу

Звук − це явище, що становить собою розповсюдження у вигляді пружних хвиль механічних коливань у твердому, рідкому або газоподібному середовищі. У фізичному сенсі під звуком мають на увазі коливання, що класифікуються в зв'язку з тим, як вони будуть сприйняті органами чуття людини. Як і всі хвилі, звук має три характеристики, а саме: амплітуду, період та час (Рис. 1.1). Людина здатна чути звукові коливання в діапазоні частот від 15-21 Гц до 14-21 кГц. Звук який знаходиться нижче діапазону чутності людини називають інфразвуком; вище: до 1 ГГц, − ультразвуком, від 1 ГГц і вище − гіперзвуком. Гучність звуку, так чи інакше, залежить від ефективного звукового тиску, форми і частоти коливань, а висота звуку – не тільки від величини звукового тиску, а й від частоти [4].

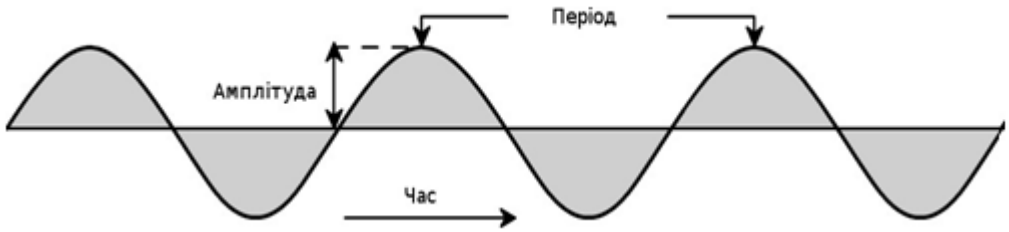

Рисунок 1.1 − Схематичне зображення звукової хвилі.

Серед звуків, які може сприймати людина, слід виділити музичні звуки (з яких складається музика) й фонетичні, мовні звуки і фонеми (з яких складається усне мовлення). Музичні звуки містять кілька тонів, а іноді і шумові компоненти в дуже широкому діапазоні частот. Надалі будуть розглянуті саме мовні звуки.

У комп'ютерній техніці звук зберігається у цифровому вигляді. Цифровий звук − результат трансформації аналогового звукового діапазону в цифровий формат. Існує декілька алгоритмів трансформації звукового сигналу у цифровий аудіо формат. Найпростіший алгоритм трансформації, імпульсно-кодова модуляція (ІКМ), полягає в поданні послідовних значень рівня сигналу, що перетворюються аналого-цифровим перетворювачем (АЦП) через рівні проміжки часу (Рис. 1.2).

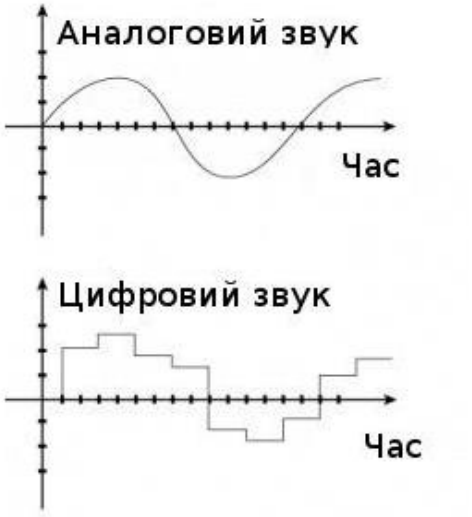

Рисунок 1.2 − Порівняння аналогового сигналу та цифрового сигналу.

Але такий формат інформації не є задовільним для розпізнавання тому, що цифровий сигнал все ще представлений з накладенніми одна на одну хвилями, нехай і в цифровому аудіо форматі (рис. 1.3).

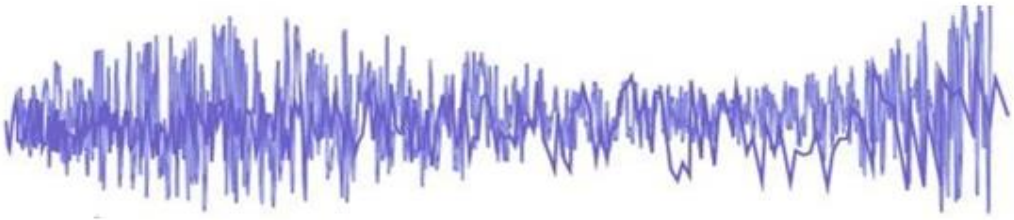

Рисунок 1.3 − Накладання сигналів, що утворюють звукову хвилю.

Основні характеристичні риси мови людини, пов'язані з розмірами, формою, динамічними змінами мовного тракту і показують емоційний стан людини. Їх можна поділити на чотири класифікаційні групи ознак, що дозволяють знаходити відмінності в мовних зразках: спектрально-часові, кепстральні, амплітудно-частотні та ознаки нелінійної динаміки.

Спектральні ознаки:

- усереднєне значення спектра звукового сигналу;
- нормалізація середнього значення спектру;
- час знаходження звукового сигналу в спектральних смугах;
- нормальний час знаходження мовного сигналу в смугах спектру;
- середне значення мовного спектра в смугах;
- повна сила мовного спектра в смугах.

Часові ознаки:

- часовий проміжок сегмента;
- висотність сегмента;
- числове значення форми сегмента.

Спектрально-часові характеристики визначають мовний сигнал в його

фізико-математичній суті завдяки наявності елементів трьох видів:

- дискретних (тональних) ділянок мовної хвилі;

- статичних ділянок мовного сигналу (шумових);

- ділянки без мовних пауз.

Спектрально-часові характеристики відображають індивідуальність форми дискретного ряду і спектру мовних імпульсів у різних людей і особливі фільтруючі функції їх мовлення. Характерними особливостями звукового потоку, пов'язаними з динамікою перебудови артикуляційних органів мови і є інтегральні характеристики звукового потоку, що відображають за допомогою взаємозв'язку або одинаковості руху артикуляційних органів людини.

Кепстральні ознаки:

- коефіцієнти рівного передбачення із вчитуванням змін на дискретність чутливості вуха мовця;

- коефіцієнти сили частоти створення;

- коефіцієнти спектра рівного прогнозування;

- коефіцієнти кепстра рівного прогнозування.

Більшість нинішніх комп’ютерних систем розпізнавання мови зосереджують ресурси на отриманні частотної характеристики мовного тракту мовця, не зважаючи при цьому на характеристики сигналу збудження. Це зв’язано з тим, що коефіцієнти першої моделі гарантують кращу чіткість звуків [5].

Однією з головних проблем при достигненні цілі трансформації вхідної звукової хвилі у цифровий аудіо формат є надання малого значення фоновому шуму довкілля.

Шум або акусти́чний шум − коливання частинок довкілля, що сприймається органами слуху людини як небажані сигнали. З точки зору акустики: шум − нестійкі або випадкові акустичні коливання, що характеризуються випадковою зміною амплітуди і частоти.

Шуми можна класифікувати :

а) спектральні:

1) статичні шуми;

2) динамічні шуми.

б) за характером:

1) широкосмуговий шум зі статичним спектром завширшки одну октаву;

2) тональний шум, в спектрі якого є виражені тони.

в) за частотою:

1) з низькою частотою (<310 Гц);

2) з середньою частотою (310-810 Гц);

3) з високою частотою (>810 Гц).

г) за часом:

1) непостійні;

2) постійні.

ґ) за фізичною виникнення:

1) гідравлічний;

2) аеродинамічний;

3) механічний;

4) електромагнітний.

Відношення сигнал-шум (ВСШ; англ. Signal-to-noise ratio, скор. SNR)− безрозмірна величина, що дорівнює відношенню потужності корисного сигналу до потужності шуму. Можна описати формулою:

$$\mathrm{SNR} = \frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}} = \left(\frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}}\right)^2 \qquad (1.1)$$

де P − середня потужність, а A − середньоквадратичне значення амплітуди. Обидва сигнали вимірюються в смузі пропускання системи.

Зазвичай відношення сигнал-шум виражається в децибелах (дБ). Чим більше це відношення, тим менше шум впливає на характеристики системи.

$$\mathrm{SNR(dB)} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) = 20\log_{10}\left(\frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}}\right) \qquad (1.2)$$

Основними причинами великого шуму в звукових сигналах є:

- порушення каналів передачі звукового сигналу;
- тепловий шум і дробовий шум в компонентах системи;
- недостатня розрядність АЦП;
- резонансні явища;
- паразитні зв'язки (паразитна ємність);
- самозбудження системи;
- дискретність передавальних характеристик.

Шум квантування виникає при перетворенні аналогового сигналу в цифровий. Тоді коли аналоговий сигнал − безперервний і в ідеалі може мати нескінченну точність, точність цифрового сигналу залежить від частоти квантування та бітової розрядності аналого-цифрового перетворювача.

Різниця між вихідним аналоговим сигналом та оцифрованим обумовлена «округленнями», які іменуються терміном похибки квантування (рис. 1.4).

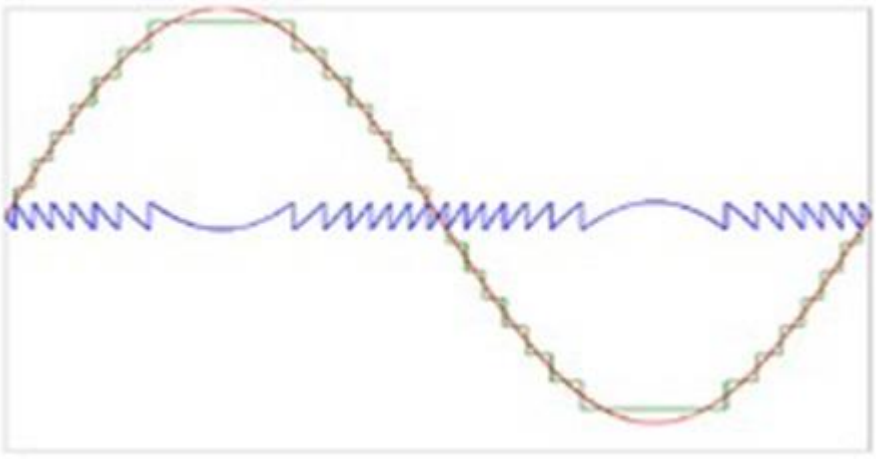

Рисунок 1.4 − Помилка квантування (синя лінія) синусоїдального сигналу (червона лінія). Квантований сигнал − зелена лінія.

Шум квантування можна подати як адитивний уривчастий сигнал *e(nT)*, що враховує похибки квантування. Якщо *d(nT)* – вхідний звуковий сигнал квантувального апарата, а *F[]* − його функція що передає, то з цього випливає наступна рівна модель шуму квантування:

$$e(nT) = F[d(nT)] - d(nT) \tag{1.3}$$

Рівна модель використовується для аналітичного дослідження властивостей шуму квантування.

Зниження шуму − процес очищення корисного сигналу від шумів з метою підвищення його якості або для зменшення рівня помилок у лініях передачі і системах зберігання цифрових даних. Методи зниження шуму концептуально дуже схожі незалежно від оброблюваного сигналу, проте попереднє знання характеристик сигналу, що передається може значно вплинути на реалізацію цих методів в залежності від типу сигналу.

Системи зниження шуму (СЗШ) − системи обробки сигналу, реалізовані у вигляді електронних схем або програмних алгоритмів, призначені для збільшення ВСШ за рахунок надмірності або зниження розрядності або дозволу сигналу. Також для позначення СЗШ часто застосовується термін «шумознижувач».

Системи шумозниження широко використовуються як для обробки звукового (аудіо) сигналу, так і для відео (фото) сигналу. Більшість СЗШ поділяються на два типи:

- фільтрація. СЗШ обробляє сигнал при прийомі (відтворенні), або запису (передачі) намагаючись очистити корисний сигнал від шуму;
- системи, що модифікують сигнал для передачі по шумним каналам (або для запису сигналу на носій), з наступним оберненим перетворенням на приймальній стороні (при відтворенні).

В системах розпізнавання голосових сигналів найчастіше використовується перший тип СЗШ [7].

В даній роботі часто буде вживатися термін «великий словник», або «обмежений словник». Залежно від контексту числове представлення даних

термінів може змінюватися. На приклад, якщо будувати систему розпізнавання голосових сигналів орієнтуючись на обчислювальні потужності сучасних персональних комп'ютерів, то числове представлення терміну «великий словник» може становити кількість слів всіх сучасних популярних мов. На противагу, якщо брати в якості орієнтиру потужності будь-якого побутового пристрою, числове представлення терміну «великий словник» може становити 100 слів.

1.2.2 Задача аналізу параметризованого звукового сигналу

Після перетворення вхідного звукового сигналу, отримаємо набір параметрів, тобто числове представлення окремого фрейму. Тепер постає задача безпосереднього розпізнавання вхідного сигналу. Іншими словами, постає задача перетворення набору цифрових коефіцієнтів у слово. Залежно від задачі системи розпізнавання та алгоритму, що використовується, задача перетворення цифрового сигналу може змінитися на задачу порівняння вхідного сигналу з деяким, вже існуючим, еталоном [8].

На стан вхідного сигналу впливають такі характеристики, як швидкість мовлення, гучність, наявність шумів, тембр голосу диктора, чіткість вимови слів. Тобто одне слово може характеризуватися великою кількістю різних сигналів. Це накладає велику кількість обмежень на використання алгоритмів для розпізнавання, адже при обробці голосових сигналів необхідно враховувати їх сильну мінливість.

Отже, алгоритми, що базуються на порівнянні з еталоном придатні для використання тільки при сильно обмеженому словника. У свою чергу, алгоритми, що базуються на перетворенні є складними у реалізації, часто потребують великих обчислювальних можливостей та не гарантують ідеальну точність розпізнавання [9].

Підхід, що базується на порівнянні вхідного сигналу з еталонами був першим, який досить успішно вирішував задачу розпізнавання голосових

сигналів. Але з розвитком технологій розпізнавання з’являлися нові алгоритми, які використовували інші підходи до процесу розпізнавання і не потребували еталонів для свого функціонування. Наступним успішним алгоритмом, що використовувався для вирішення задач розпізнавання голосових сигналів став алгоритм, що використовує приховані марковські моделі, які базуються на описі стохастичних процесів та статистичних законах. В наш час найбільш популярні системи розпізнавання використовують нейронні мережі, адаптовані до задач розпізнавання голосових сигналів [10].

1.2.3 Задача розпізнавання голосових сигналів за допомогою алгоритму динамічної трансформації часової шкали

Часові ряди − широко поширене представлення даних, зустрічається, фактично, в будь-якої наукової області, і порівняння двох послідовностей є стандартною задачею. Для обчислення відхилення буває досить простого вимірювання відстані між компонентами двох послідовностей (евклідова відстань). Однак часто дві послідовності мають приблизно однакові загальні форми, але ці форми не вирівняні по осі X. Щоб визначити подобу між такими послідовностями, ми повинні «деформувати» вісь часу однієї (або обох) послідовностей, щоб досягти кращого вирівнювання [19].

Вимірювання відстані між двома часовими рядами потрібно для того, щоб визначити їх подібність і класифікацію. Таким ефективним видом виміру є Евклідова метрика. Для двох часових послідовностей це просто сума квадратів відстаней від кожної n-ої точки однієї послідовності до n-ої точки іншої послідовності. Однак використання Евклідової відстані має істотний недолік: якщо два часових ряди однакові, але один з них незначно зміщений у часі (уздовж осі часу), то Евклідова метрика може порахувати, що ряди відрізняються один від одного.

1.3 Аналіз існуючих рішень розпізнавання голосових сигналів

Обробка голосового сигналу починається з його оцифровки. Для цього необхідно заздалегідь записати його в оперативну пам'ять комп'ютера або на машинний носій. Як було сказано вище, більшість персональних комп'ютерів вже оснащені обладнанням, необхідним для введення і виведення звуку. Це мікрофон і звукова плата. У загальному вигляді процес введення мовних повідомлень.

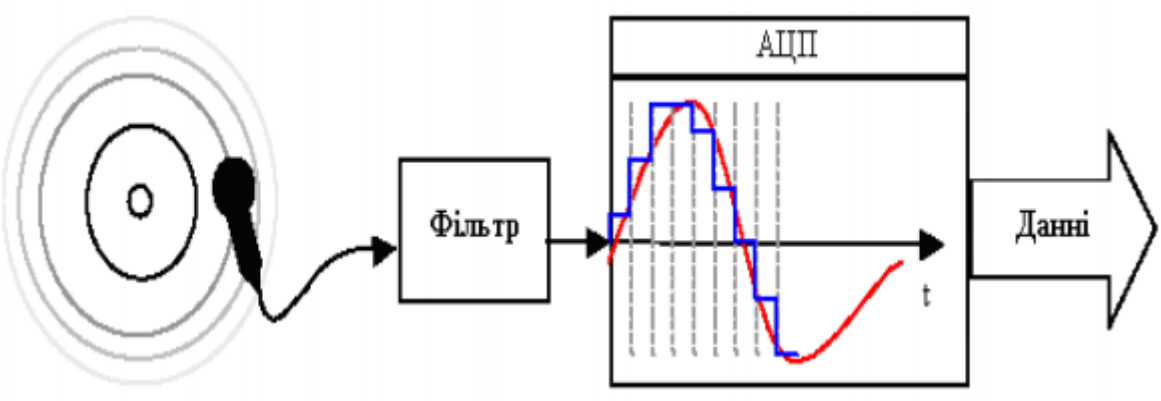

Рисунок 1.5 – Схема вводу голосових повідомлень в персональний комп’ютер

Голосовий сигнал формується і передається в просторі у вигляді звукових хвиль. Джерелом голосового сигналу є мовоутворюючий тракт людини. Приймачем сигналу є датчик звукових коливань, мікрофон − пристрій перетворення звукових коливань в електричні. Існує велика кількість типів мікрофонів (вугільні, електродинамічні, електростатичні, п'єзоелектричні та ін. Деякі з цих мікрофонів для своєї роботи вимагають зовнішнього джерела струму (наприклад, вугільні та конденсаторні), інші під впливом звукових коливань здатні самостійно виробляти змінна електрична напруга (це електродинамічні і електретні мікрофони). Є також мікрофони, призначені спеціально для комп'ютерів. Такі мікрофони зазвичай кріпляться на підставці, що стоїть на поверхні столу. Комп'ютерні мікрофони можуть комбінуватися з

головними телефонами.

Як же вибрати з усього різноманіття мікрофонів той, що найкраще підходить для систем розпізнавання мовлення? В принципі, можна експериментувати з будь-яким наявним мікрофоном, якщо тільки його можна підключити до звукового адаптера комп'ютера. Однак розробники систем розпізнавання мовлення рекомендують придбати такий мікрофон, який при роботі буде перебувати на постійному відстані від рота диктора.

Якщо відстань між мікрофоном і ротом не змінюється, то середній рівень електричного сигналу, що надходить від мікрофона, також буде мінятися не занадто сильно. Це матиме позитивний вплив на якість роботи сучасних систем розпізнавання мови. Якщо мікрофон стоїть на столі, то при повороті голови або зміні положення тіла відстань між ротом і мікрофоном буде змінюватися. Це призведе до зміни рівня вихідного сигналу мікрофона, що, в свою чергу, погіршить надійність розпізнавання мовлення. Тому при роботі з системами розпізнавання мовлення найкращі результати будуть досягнуті, якщо використовувати мікрофон, прикріплений до головних телефонів. При використанні такого мікрофона відстань між ротом і мікрофоном буде постійним.

Чутливим елементом мікрофона будь-якого типу є пружна мембрана, яка залучається до коливальний процес під впливом звукових хвиль. Мембрана пов'язана з перетворюючим елементом, який перетворює коливання мембрани в електричний сигнал.

З виходу мікрофону сигнал подається на вхід звукової карти персонального комп'ютера. Величина вхідного сигналу, що надходить від мікрофона, змінюється періодично і приймає як позитивні, так і негативні значення. При записі звукова карта представляє собою аналого-цифровий перетворювач з можливостями налаштування параметрів оцифровки.

Основними параметрами є частота дискретизації і розрядність кодування. Дані параметри визначають якість і розмір одержуваної вибірки в результаті запису. Вибір частоти дискретизації безпосередньо залежить від узагальненої

спектральної щільності потужності мовного сигналу (рис. 1.6). Узагальнена спектральна щільність потужності має максимум в діапазоні 250-500 Гц і затухає зі швидкістю, що дорівнює 8-10 дБ на октаву (при подвоєнні частоти). Це призводить до того, що на частотах вище 4000 Гц спектральна щільність падає до рівня 60 дБ, що відповідає послабленню потужності в порівнянні з максимумом (-25 ... -30 дБ) в 20 і більше разів. Це дозволяє вважати, що смуга пропускання для каналів передачі звукових повідомлень може бути обмежена частотою 4-5 кГц. Відповідно теореми Котельника, частота дискретизації цього сигналу повинна становити 8-10 кГц. Зазначимо, що частота дискретизації 8 кГц являється стандартною для телефонних апаратів.

В звукових картах персональних комп'ютерів ця частота є мінімальною, при цьому передбачена можливість суттєво підвищити частоту дискретизації. Таким чином, можна вважати, що звукова карта персонального комп'ютера дозволяє з достатньою якістю дискретизувати голосовий сигнал.

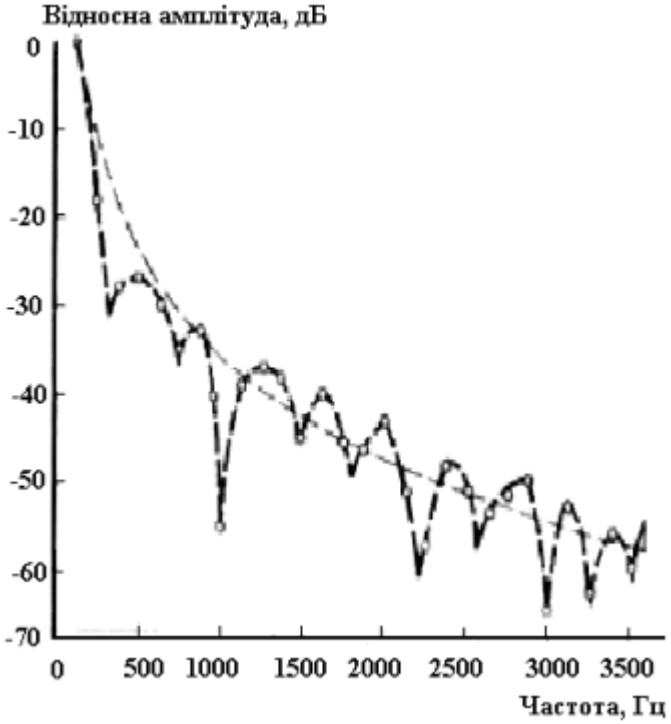

Рисунок 1.6 – Графік амплітудного спектру потужності голосового сигналу

В сучасних звукових картах використовується імпульсно-кодова модуляція, при якій кожен дискретний відлік мовного повідомлення кодується відповідно з деякими правилами. Розраховано, що для забезпечення співвідношення сигнал/шум квантування, рівного 36 дБ, потрібно не менше семи двійкових розрядів і що для отримання високоякісного цифрового кодування сигналу мовлення необхідно 11 розрядів. На практиці кількість розрядів визначається розрядністю персонального комп'ютера. Як правило, вона дорівнює або кратна восьми розрядам. При використанні найбільш поширеного програмного забезпечення кількість розрядів дорівнює 16, 32 або 64. Першочергова фільтрація шумів в дискретизованому сигналі полягає в накладенні на цей сигнал вікон різного типу – Кайзера, Хемінга та інших. В доступній літературі не знайдено критеріїв, по яким вибирається тип вікна.

Після фільтрації шумів для виділення із звукового потоку окремих слів застосовується аналіз енергії сигналу на протязі кожних 10-20 мс. Крім того, визначити початок/кінець слова можна по всплеску/затуханню величини сигналу. Обробка вхідного оцифрованого сигналу з метою зменшення обсягу вхідних даних полягає у застосуванні різних методів спектрального аналізу даних. Спектральний аналіз даних реалізується або за допомогою віконного дискретного перетворення Фур'є або за допомогою дискретних вейвлет-перетворень. Зазначимо, що діапазон частот, які чує людина, знаходиться в межах від 16 Гц до 20000 Гц. Однак в більшості систем, виходячи в тому числі і із можливостей комп'ютерної звукової апаратури, застосовується діапазон частот від 50 Гц до 16000 Гц. На сьогодні загальноприйнято проводити стиснення спектру за допомогою процедури визначення мел-кепстральних коефіцієнтів. Її результатом являється 16-24 коефіцієнтів, які достатньо повно характеризують весь діапазон звукових частот, які відчуває людина. Ще один підхід стиснення спектру базується на методі визначення формант голосового сигналу [7, 8].

Для порівняння еталонного та піддослідного сигналів в теперішній час в основному використовуються сховані марківські моделі, методи динамічного

програмування та нейронні мережі.

Використання схованих марківських моделей базується на постулаті, що голосовий сигнал може бути розділений на стаціонарні фрагменти, які відповідають окремим станам ланцюга Маркова

$$O \{o ,o ,...,o \}. = 1\ 2\ \tau \qquad (1.4)$$

Перехід між станами відбувається миттєво, а ймовірність відображення породженого моделлю фрагменту залежить тільки від поточного стану моделі та не залежить під попередніх станів.

Власне схованою марківською моделлю називається модель, що складається із N станів, в кожному із яких деяка система може приймати одне з M значень деякого параметра. Як правило, модель задається виразом.

$$\lambda = \{A,B,\pi\}, \qquad (1.5)$$

де A – матриця ймовірностей переходів по станам, B – вектор ймовірностей випадіння кожного із M значень параметру в кожному із N станів, π – вектор розподілу початкових ймовірностей. Перший етап використання моделі полягає у її навчанні на прикладах, що відповідають еталону ключового слова.

Результатом навчання являється розрахунок параметрів виразу (1.5). Після цього на вхід моделі можна подавати послідовність, яка відповідає невідомому голосовому сигналу. Вирішення задачі знаходження ймовірності появи цієї послідовності у кожній із попередньо навчених моделей дозволить визначити ту модель, яка найбільш достовірно відповідає голосовому сигналу, а значить, і розпізнати ключове слово. До основних недоліків схованих марківських моделей відносять велику обчислювальну складність та складність формування бази даних ключових слів. Застосування методів динамічного програмування зводиться до розрахунку ключового слова найбільш схожого на невідоме.

Критеріями схожості слів виступають відстань Евкліда, відстань Хемінга та інші. В доступній літературі методики вибору критерію схожості не знайдено.

Використання нейронних мереж базується на їх здатності класифікувати голосові сигнали, задані за допомогою коефіцієнтів, які відповідають спектральним характеристикам [10]. Не зважаючи на перспективність даного напрямку, застосуванню нейронних мереж перешкоджає відсутність методики оптимізації типу та параметрів мережі. В підсумку можна зазначити, що основною невирішеною задачею в області пошуку ключових слів є задача порівняння еталонних та невідомих голосових фрагментів.

На сьогодні перераховані технології розпізнавання мовних сигналів реалізовані в наступних програмних комплексах:

- програми голосового управління комп'ютером VoiceNavigator, Truffaldino;
- бібліотека розпізнавання голосових команд VoiceCom;
- систему голосового розмежування доступу, розроблену компанією «Центр мовних технологій»;
- програми документування усних виступів - комп'ютерний транскрайбер, системи Нестор і Алегро.

Програма VoiceNavigator – є типовим представником програм голосового управління комп'ютером. Вона дозволяє користувачеві запускати додатки голосом, не торкаючись клавіатури, і виконувати довільно задані команди. Перед використанням програми VoiceNavigator її необхідно навчити, вимовивши в мікрофон слова команд. Так як програма VoiceNavigator розпізнає команди за зразками, то команди можна вимовляти будь-якою мовою і будь-яким голосом.

Щоб програма почала розпізнавати голосові команди, її необхідно «розбудити», вимовивши ключове слово. Після цього програма буде реагувати тільки на ваші команди, ігноруючи інші звуки. У програмі є функція голосової відповіді-підтвердження команд. Ця функція дозволяє переконатися, що ваша команда розпізнана системою і готова для виконання. Програма VoiceNavigator невимоглива до ресурсів комп'ютера. Ви можете використовувати її в комп'ютері, обладнаному процесором з тактовою

частотою 200 МГц або вище, причому для введення звукових команд підійде будь-який звуковий адаптер, наприклад, Creative Sound Blaster.

Бібліотека розпізнавання голосових команд VoiceCom. – становить ядро описаних вище програм VoiceNavigator і Truffaldino. З її допомогою розробники можуть додати голосове управління в створювані ними програми. Скориставшись готової бібліотекою VoiceCom, розробники можуть легко додати у додатки наступні функціональні можливості: управління обладнанням за допомогою голосу; виконання мовних запитів до баз даних через мікрофон або навіть по телефону; пошук за ключовими словами в звукових WAV-файлах. Слід зазначити, що бібліотека VoiceCom дозволяє вбудовувати голосові функції не тільки в звичайні програми для персональних комп'ютерів, але і в автономні пристрої, обладнані цифровими сигнальними процесорами DSP. Алгоритми, реалізовані в бібліотеці розпізнавання голосових команд VoiceCom, мають високу швидкодію, невибагливі до обсягу оперативної пам'яті і здатні адаптуватися до шумів. Бібліотека VoiceCom забезпечує розпізнавання команд, виголошених будь-яким голосом і будь-якою мовою. При цьому є можливість структурування для практично необмеженого словника. При цьому алгоритми дозволяють розпізнавати 100-200 команд з попереднім навчанням для кожного диктора, і 30-50 команд для будь-якого диктора (в режимі, не залежному від диктора). Якщо команди вимовляються по телефону, то алгоритми бібліотеки VoiceCom дозволяють розпізнати 10-20 слів, вимовлених яким диктором. Ну і, звичайно, у бібліотеці реалізована можливість активації розпізнавання команд за ключовим словом, що виключає несподівані реакції системи на сторонні звуки.

Голосове розмежування доступу за допомогою бібліотеки VoiceKey Kit. Компанія «Центр мовних технологій» створила бібліотеку розмежування доступу по голосу VoiceKey Kit, яку можна легко вмонтувати в будь-які додатки. Це можуть бути офісні додатки, комп'ютерні ігри, системи «батьківського контролю» та ін. Ця бібліотека дозволяє розпізнавати парольні фрази (типу «Сезам, відкрийся!»), або особливості голосу тієї чи іншої

людини. В якості її недоліків вказують низьку надійність розпізнавання.

Системи Нестор і Алегро. Система Нестор призначена для багатоканальної цифрової звукозапису та оперативного текстового розшифрування кількох усних виступів і фонограм мови за принципом розподіленої обробки (стенографування). Комплекс Нестор забезпечує синхронну обробку до 24 акустичних каналів (виступаючих і/або фонограм мови). В цю систему входить комп'ютер станції звукозапису, обладнаний 4-каналоним звуковим адаптером і спеціалізованим програмним забезпеченням і звуковий сервер для архівування звукових записів на диски CD-RW. Комплекс Нестор може комплектуватися педаллю для управління відтворенням звукового сигналу. В системі передбачені робочі місця адміністратора, керівника групи та операторів. В залежності від варіантів поставки в комплексі може бути від 3 до 50 робочих місць операторів, від 1 до 8 робочих місць керівників груп і 1-2 робочих місця адміністратора. Таким чином, система Нестор придатна для автоматизованого документування досить великих нарад і форумів. Поступаючий на її вхід мовної сигнал записується на жорсткий диск комп'ютера. Потім він розбивається на фрагменти і розподіляється між операторами-стенографістами, які виконують його прослуховування і розшифрування. Отримані таким чином ділянки тексту автоматично з’єднуються в єдиний документ, який після перевірки може бути збережений і роздрукований.

Програми для диктування тексту. Сьогодні існують потужні програми, здатні не тільки розпізнавати і виконувати команди, а й розпізнавати мову в режимі диктування. Як правило, такі програми або забезпечуються власним редактором тексту, або здатні працювати з будьякими редакторами тексту і таблиць, наприклад, такими, як Microsoft Word і Microsoft Excel. Крім того, система для диктування тексту входить до складу браузеру Chrome та сучасних версій операційної системи Android. Основним недоліком вказаних систем є високі обчислювальна ресурсоємність.

1.4 Висновки до розділу 1

В першому розділі проведено аналіз предметної області «ефективність методів розпізнавання мови ».

Також в розділі було зроблено постановку мети та завдань дослідження, поставлено задачу та розглянуто базові поняття про роспізнавання мови, було розглянуто алгоритми, які використовуються на сьогоднішній момент.

Розглянуто основні проблеми розпізнавання мови та визначено задачі які ставить перед собой алгоритми розпізнавання мови

Були розглянуто існуючі реалізовані системи та проаналізовано принцип роботи олгоритмів розпізнавання мови.

2 ПЛАНУВАННЯ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ ЕФЕКТИВНОСТІ МЕТОДІВ РОЗПІЗНАВАННЯ МОВИ

2.1 Методи, які застосовуються для розпізнавання мови

Сьогодні системи розпізнавання мови будуються на основі принципів визнання форм розпізнавання. Методи і алгоритми, які використовувалися до сих пір, можуть бути розділені на наступні великі класи: Класифікація методів розпізнавання мови на основі порівняння з еталоном.

- динамічне програмування;
- тимчасові динамічні алгоритми (Dynamic Time Warping).

Контекстно-залежна класифікація. При її реалізації з потоку мови виділяються окремі лексичні елементи:

- фонеми і Алофон, які потім об'єднуються в склади і морфеми;
- приховані Марківські моделі (Hidden Markov Model);
- нейронні мережі (Neural networks).

2.1.1 Застосування DWT алгоритму в розпізнаванні мови

Звук проходить через середовище, як поздовжня хвиля зі швидкістю, яка залежить від щільності середовища. Найпростіший спосіб представлення звуків − синусоїдний графік. Графічне представлення вібрацій повітря під тиском протягом деякого часу.

Форма звукової хвилі залежить від трьох чинників: амплітуди, частоти і фази. Амплітуда − переміщення синусоїдальних графів вище і нижче тимчасової осі (y = 0), що відповідає енергії завантаженої звукової хвилі. Вимірювання амплітуди може бути вироблено в одиницях тиску (децибелах DB), які вимірюють амплітуду звичайного звуку за допомогою логарифмічних функцій. Вимірювання амплітуди використовуючи децибели дуже важливо на практиці, так як це пряме уявлення про те, як гучність звуку сприймається

людьми. Частота − число циклів сінусоїди за одну секунду. Цикл коливань починається з середньою лінією, потім досягає максимуму і мінімуму, а після повертається до середньої лінії. Частота циклу вимірюється за одну секунду або в герцах (Гц).

Величина зворотна частоті називається періодом − час, необхідний звуковий хвилі для завершення циклу. Останній фактор − фаза. Вона вимірює положення щодо початку синусоїдальної кривої. Фаза не може бути почута людиною, однак її можна визначити щодо положення між двома сигналами. Проте, слуховий апарат сприймає положення звуку на різних фазах.

Для того щоб розібрати звукові хвилі на синусоїдальної кривої треба скористатися теоремою Фур'є. У ньому записано, що будь-яка комплексна періодична хвиля може бути розібрана за допомогою синусоїдальної кривої з різними частотами, амплітудами і фазами. Цей процес називається аналіз Фур'є, і його результатом є набором амплітуд, фаз і частот для кожного синусоидального компонента хвилі. Складаючи ці синусоїдальні криві разом, виходить оригінальна звукова хвиля. Точка частоти або фази, взята разом з амплітудою, називається спектром. Будь-періодичний сигнал показує, рекурсивную модель часу, яка відповідає першій частоті коливань сигналу і називається основною частотою. Вона може бути виміряна з мовного сигналу, за допомогою перевірки періоду коливань близько 0 осі. Спектр показує частоту короткої послідовності звуків, і якщо ми хочемо проаналізувати її розвиток протягом часу, необхідно знайти спосіб, що дозволяє продемонструвати це. Це можна показати на спектрограмі. Спектрограма − це діаграма в двох вимірах: частота і час, − в якій колір точки (темний − сильний, світлий − слабкий) визначає амплітуду інтенсивності. Метод грає важливу роль в розпізнаванні мови, і професіонал може розкрити багато подробиць, дивлячись тільки на звукову спектрограму.

Сучасні методи виявлення можуть точно визначити початкову та кінцеву точку сказаного слова в звуковому потоці, на основі обробки сигналів мінливих протягом часу. Дані методи оцінюють енергію і середню величину в

короткому відрізку часу, а також обчислюють середній рівень перетину нуля.

Створення початкової і кінцевої точки − проста завдань, якщо аудіозапис зроблена в ідеальних умовах. У цьому випадку відношення сигнал-шум великий, так як визначити дійсний сигнал в потоці шляхом аналізу образів не становить труднощів. В реальних умовах все не так просто: фоновий шум має величезну інтенсивність і може порушити процес відділення слів в потоці мовлення.

Кращий алгоритм відділення слів − алгоритм Рабінел-Ламель. Якщо розглядати строб - імпульсів {s1, s2, ..., sn}, де n − число образів строб-імпульсів, а si, i = 1, n − чисельне вираження зразків, загальна енергія строб-імпульсів обчислюється:

$$E(n) = \frac{1}{n}\sum_{i=1}^{n} s_i^{\,2}. \qquad (2.1)$$

Середній рівень перетину нульового рівня:

$$ZCR(n) = \sum_{i=1}^{n-1} sign(s_i) \cdot sign(s_{i+1}), \qquad (2.2)$$

де:

$$sign(s_i) = \begin{cases} 1 & \text{if } s_i > 0 \\ 0 & \text{if } s_i < 0 \end{cases}. \qquad (2.3)$$

Метод використовує три числових рівня: два для енергії (верхній, нижній) і один для середнього перетину нульового рівня. Точка, починаючи з якої енергія перекриває верхній рівень і рівень позитивних і негативних значень, не скасовує встановлений рівень, який вважається відправною точкою голосового звучання (НЕ тиші). Пошук першої такої точки проводиться шляхом схрещування імпульсів від початку і до кінця, і це визначить першу область з промовою. Зворотний перехід, з кінця в початок, дозволяє визначити кінцеву точку останньої області з промовою. Визначення

всередині області може бути зроблено шляхом схрещування імпульсів між двома цими точками. Початок глухий області починається в точці, в якій енергія стає менше значення нижнього рівня. Зверніть увагу на малюнок нижче, на якому до і після видалення глухий області.

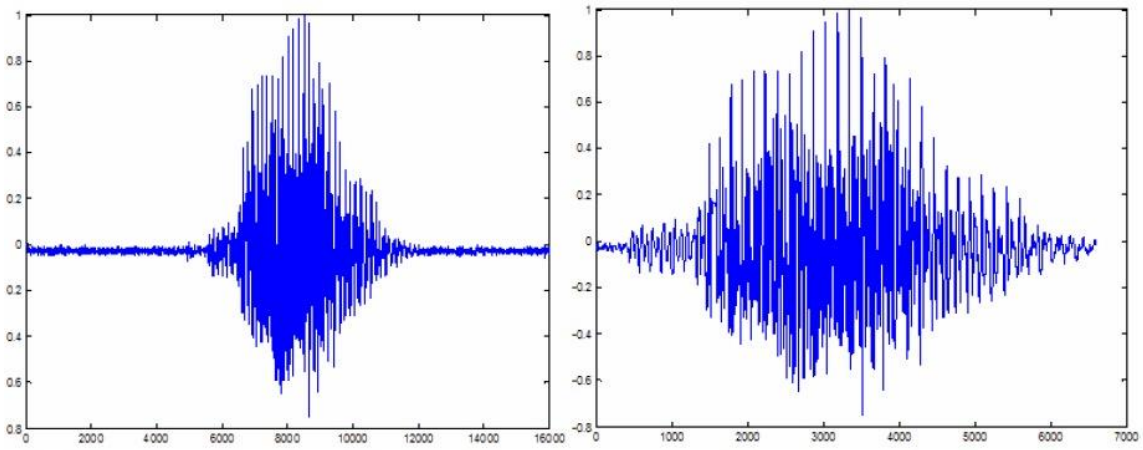

Рисунок 2.1 − Звуковий сигнал слова «no».

Визначення слова може здійснюватися шляхом порівняння числових форм сигналів або шляхом порівняння спектрограми сигналів. Процес порівняння в обох випадках повинен компенсувати різні довжини послідовності і нелінійний характер звуку. DWT алгоритму вдається розібрати ці проблеми шляхом знаходження деформації, відповідної оптимальному відстані між двома рядами різної довжини.

2.1.2 Застосування прихованих Марківських моделей для розпізнавання мови

Використання прихованих Марківських моделей для розпізнавання мови засноване на двох наближеннях:

1) може бути розбита на фрагменти, що відповідають станам в СММ, параметри мови в межах кожного фрагменту вважаються постійними.

2) ймовірність кожного фрагмента залежить тільки від поточного стану системи і не залежить від попередніх станів. Модель називається

«прихованою», так як нас, як правило, не цікавить конкретна послідовність станів, в якій перебуває система. Ми або подаємо на вхід системи послідовності типу O = {o1, o2, ... oi} − де кожне oi − значення параметра (одне з M), прийняте в i-й момент часу, а на виході очікуємо модель λ = {A, B , π} з максимальною вірогідністю генеруючу таку послідовність, − або навпаки подаємо на вхід параметри моделі і генеруємо породжуються їй послідовність. І в тому і в іншому випадку система виступає як "чорний ящик", в якому сховані дійсні стану системи, а пов'язана з нею модель заслуговує назви прихованої. Для здійснення розпізнавання на основі прихованих моделей Маркова необхідно побудувати кодову книгу, яка містить безліч еталонних наборів для характерних ознак мови (наприклад, коефіцієнтів лінійного передбачення, розподілу енергії по частотах і т.д.). Для цього записуються еталонні мовні фрагменти, розбиваються на елементарні складові (відрізки мовлення, в перебігу яких можна вважати параметри мовного сигналу постійними) і для кожного з них обчислюються значення характерних ознак. Однією елементарної складової буде відповідати один набір ознак з безлічі наборів ознак словника [1]. Фрагмент промови розбивається на відрізки, протягом яких параметри мови можна вважати постійними. Для кожного відрізка обчислюються характерні ознаки і підбирається запис кодової книги з найбільш підходящими характеристиками. Номери цих записів і утворюють послідовність спостережень O = {o1, o2, ... oi} для моделі Маркова. Кожному слову словника відповідає одна така послідовність. Далі A − матриця ймовірностей переходів з одного мінімального відрізка мови (номера запису кодової книги) в інший мінімальний відрізок мови (номер запису кодової книги). B − ймовірності випадання в кожному стані конкретного номера кодової книги .

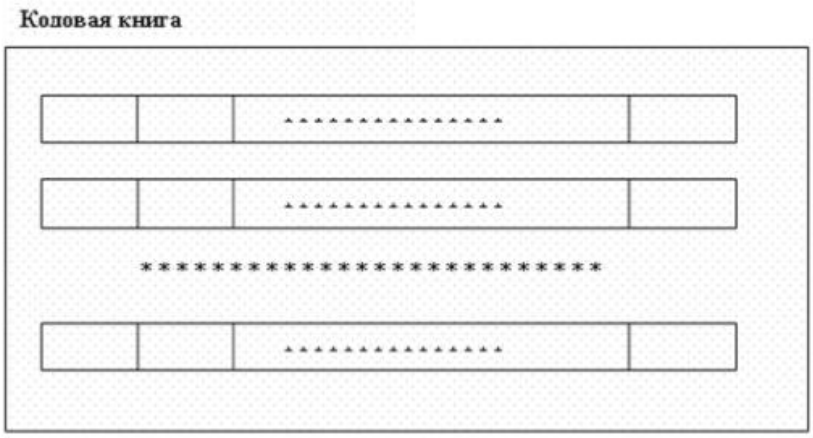

Рисунок 2.2 − Кодова книга

На єтапі налаштування моделей Маркова застосовується алгоритм Баума-Уелча для наявного словника і зіставлення кожному з його слів матриці A і B.

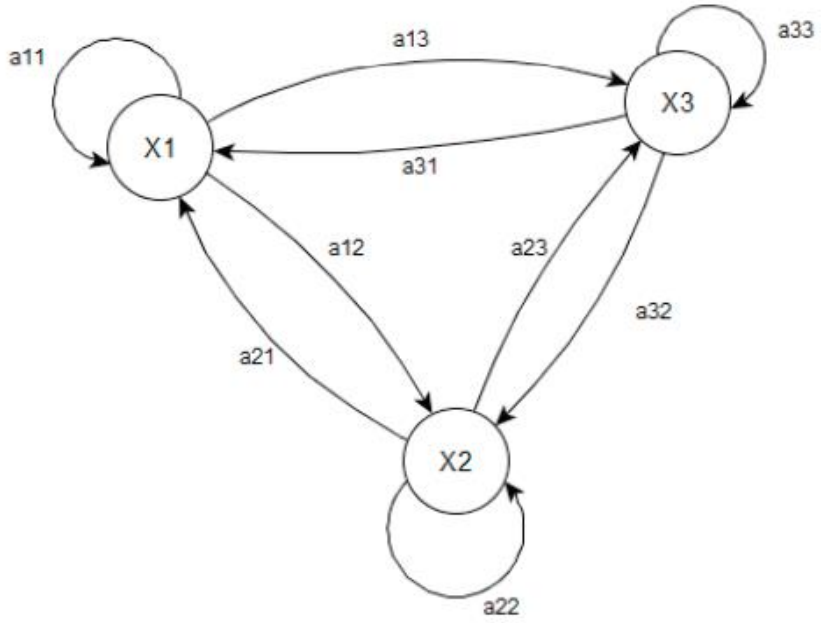

Рисунок 2.3 − Приклад діаграми переходів в прихованій марківській моделі

2.1.3 Застосування нейронних мереж для розпізнавання мови

Нейронна мережа − це математична модель, побудована на принципах роботи людського мозку. Нервова система і мозок людини складаються з нейронів, з'єднаних один з одним нервовими волокнами, які здатні передавати електричні імпульси. Нейрон складається з тіла, і відростків нервових волокон двох типів − дендритів, за якими приймаються імпульси, і одного аксона, по якому нейрон передає імпульси.

На основі нейрона був створений штучний нейрон (рис. 2.4). Його синапси представлені ваговими коефіцієнтами w1, ..., wn. Поточний стан нейрона визначається за формулою

$$S = \sum_{i=1}^{n} X_i * w_i \qquad (2.4)$$

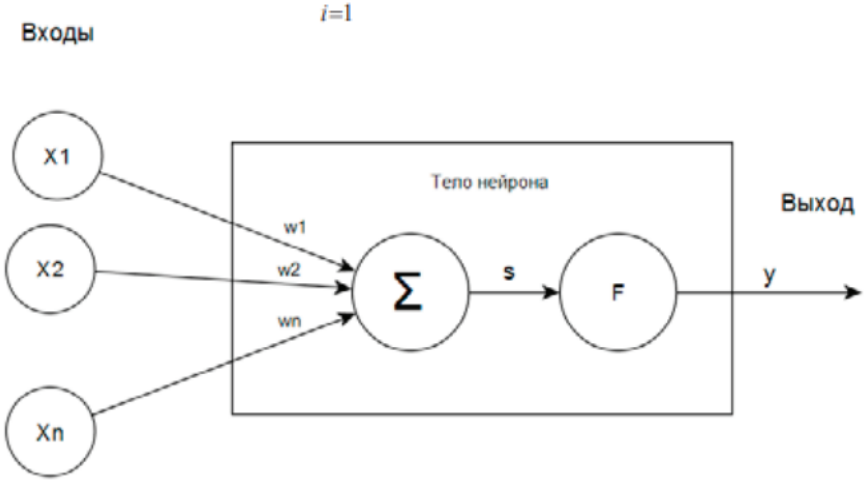

Рисунок 2.4 Схема штучного нейрона

Результат підсумовування передається в активаційну функцію F. Існують різні види активаційних функцій: порогова, лінійна, експоненціальна та інші.

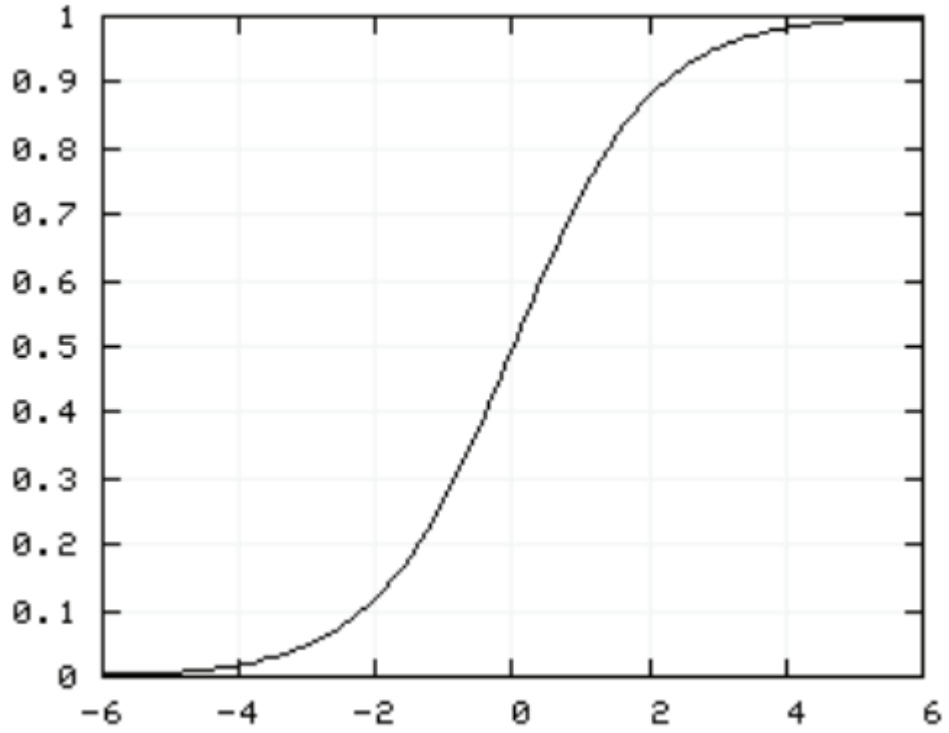

Рисунок 2.5 − Графік логістичної функції

Але найбільшого поширення набула логістична функція, яка обчислюється за формулою:

$$F(S)=\frac{1}{1+e^{-\alpha S}} \qquad (2.5)$$

де α − коефіцієнт нахилу. Змінюючи його можна будувати функції з різною крутизною.

Особливість цієї функції полягає в тому, що вона здатна підсилювати слабкі сигнали і не приводити до насичення від сильних сигналів. Нейронна мережа може складатися з одного або декількох шарів. Багатошарові мережі здатні вирішувати деякі завдання, недоступні одношаровим мереж, і володіють великими обчислювальними можливостями. Кількість входів мережі − це кількість оброблюваних ознак. Кількість виходів − це кількість можливих варіантів «відповідей» розпізнавання. Для того, щоб, подавши на

вхід нейронної мережі вектор ознак сигналу, отримати від неї близький до дійсності відповідь, її потрібно попередньо навчити. Навчання здійснюється шляхом послідовної подачі векторів ознак і одночасно з цим зміною вагових коефіцієнтів. Порядок зміни визначається обраним алгоритмом навчання. Навчання буває двох видів: з учителем і без вчителя. Навчання з учителем преполагает, що для кожного входу вектора нам відомий заздалегідь вихідний вектор, тобто навчальна вибірка складається з пар вхідних і вихідних векторів. Помилка розпізнавання, яка вичіляется як різниця виходу мережі і відомого вихідного вектора. Навчання складається в мінімізації цієї помилки і завершується, коли вона досягне прийнятного значення. Навчання без вчителя передбачає, що відповідей для вхідних векторів немає. Навчальний алгоритм повинен змінити ваги так, щоб схожі вхідні вектори, подані мережі, давали подібні відповіді. яка вичіляется як різниця виходу мережі і відомого вихідного вектора. Навчання складається в мінімізації цієї помилки і завершується, коли вона досягне прийнятного значення. Навчання без вчителя передбачає, що відповідей для вхідних векторів немає. Навчальний алгоритм повинен змінити ваги так, щоб схожі вхідні вектори, подані мережі, давали подібні відповіді. яка вичіляется як різниця виходу мережі і відомого вихідного вектора. Навчання складається в мінімізації цієї помилки і завершується, коли вона досягне прийнятного значення. Навчання без вчителя передбачає, що відповідей для вхідних векторів немає. Навчальний алгоритм повинен змінити ваги так, щоб схожі вхідні вектори, подані мережі, давали подібні відповіді.

2.2 Порівняльна характеристика існуючих алгоритмів та систем розпізнавання голосових сигналів

Технології та алгоритми розпізнавання можуть сильно різнитися залежно від поставленої задачі.

Системи розпізнавання мови можна класифікувати:

- за розміром словника (обмежений набір слів, словник великого

розміру);

- за ознакою залежності від диктора (дикторозалежні і дикторонезалежної системи);

- за типом мовлення (злита або роздільна мова);

- за призначенням (системи диктування, командні системи);

- за типом структурної одиниці (фрази, слова, фонеми, діфони, алофони);

- за принципом виділення структурних одиниць (розпізнавання за шаблоном, виділення лексичних елементів).

Існує певна кількість алгоритмів і методів для розпізнавання параметризованого звукового сигналу [11]. Залежно від основного призначення системи розпізнавання доцільно використовувати той, чи інший алгоритм. Найпопулярнішими методами для розпізнавання є:

- використання алгоритму динамічної трансформації часової шкали;

- використання прихованих марковських моделей;

- використання нейронних мереж.

Алгоритм динамічної трансформації часової шкали (DTW − dynamic time warping) є одним з алгоритмів вимірювання подібності між двома тимчасовими послідовностями, які можуть різнитися за швидкістю. Наприклад, подібність ходьби може бути виявлена за допомогою DTW, навіть якщо одна людина йшла швидше, ніж інша, або якщо під час спостереження відбувалися прискорення та сповільнення. Загалом, DTW − це метод, який обчислює оптимальну відповідність між двома даними послідовностями (наприклад, часовими рядками) з певними обмеженнями. DTW використовується щоб визначити міру подібності послідовностей, які «деформовані» нелінійно в вимірі часу, незалежно від певних нелінійних варіацій у вимірі часу. Цей метод вирівнювання послідовностей часто використовується для класифікації часових рядів [12].

На додаток до міри подібності між двома послідовностями, можна відслідкувати так званий «шлях деформації», деформуючи відповідно до цього шляху, два сигнали можуть бути вирівняні за часом. Сигнал з оригінальним

набором точок X (оригінал), Y (оригінал) перетворюється в X (деформований), Y (оригінал). Це знаходить застосування в генетичній послідовності та аудіо синхронізації. У суміжній техніці послідовності різної швидкості можуть бути усереднені за допомогою цієї техніки.

DTW має кілька слабких сторін. По-перше, складність O(n^2) не дозволяє його використання для аналізу великої кількості даних. Алгоритм погано працює при наявності шумів в сигналі, адже він немає ніяких вбудованих засобів для протидії шуму. Проте, DTW залишається простим в реалізації алгоритмом, відкритим до модифікацій. Алгоритм найчастіше використовується в додатках, яким потрібна просте розпізнавання слів: телефони, автомобільні комп'ютери, системи безпеки і подібних [13].

Штучна нейронна мережа (ШНМ) − математична модель, а також її програмне або апаратне втілення, побудоване за принципом організації та функціонування біологічних нейронних мереж − мереж нервових клітин живого організму [14].

ШНМ являє собою систему з'єднаних і взаємодіючих між собою простих процесорів (штучних нейронів). Кожний процесор подібної мережі має справу тільки з сигналами, які він періодично отримує, і сигналами, які він періодично надсилає іншим процесорам. І, тим не менше, об'єднані в достатньо велику мережу з керованою взаємодією, такі по окремо взяті прості процесори разом здатні виконати досить складні завдання.

Нейронні мережі не програмуються у звичайному сенсі цього слова, вони навчаються. Можливість навчання − одна з головних переваг нейронних мереж перед традиційними алгоритмами. Технічне навчання полягає в наведенні коефіцієнтів зв'язку між нейронами. У процесі навчання нейронна мережа здатна виявляти складні залежності між вхідними даними та вихідними. Це означає, що в разі успішного навчання мережа зможе повернути правильний результат на підставі даних, які були відсутні в початковій вибірці, а також неповних, або частково викривлених даних [15].

Прихованою Марковською моделлю (ПММ) називають модель, що

складається з N станів, в кожному з яких деяка система може приймати одне з M значень будь-якого параметра. Ймовірності переходів між станами задається матрицею ймовірностей A = {aij}, де aij − ймовірність переходу з i-го в j-ий стан. Ймовірності випадання кожного з M значень параметра в кожному з N станів задається вектором B = {bj(k)}, де bj(k) − ймовірність випадання k-го значення параметра в j-му стані. Імовірність настання початкового стану задається вектором π = {πi}, де πi − ймовірність того, що в початковий момент система опиниться в i-му стані. Таким чином, прихованою марковською моделлю називається трійка λ = {A, B, π}. Використання прихованих марковських моделей для розпізнавання мови засноване на двох наближеннях:

- може бути розбита на фрагменти, що відповідають станам в ПММ, параметри мови в межах кожного фрагменту вважаються постійними;
- імовірність кожного фрагмента залежить тільки від поточного стану системи і не залежить від попередніх станів [16].

Модель називається «прихованою», оскільки, як правило, не важлива конкретна послідовність станів, в якій перебуває система. Треба або подти на вхід системи послідовності типу O = {o1, o2, ... o?} − де кожне oi − значення параметра (одне з M), прийняте в i-й момент часу, а на виході очікувати модель ? = {A, B, ?} з максимальною вірогідністю генеруючу таку послідовність, − або навпаки подати на вхід параметри моделі і генерувати послідовність, яка продовжує її. І в тому, і в іншому випадку система виступає як «чорний ящик», в якому приховані дійсні стани системи.

Можна виділити наступні переваги використання прихованих Марковських моделей при використанні в задачі розпізнавання мови:

- HMM мають просту математичну структуру;
- структура HMM дозволяє моделювати складний ланцюжок спостережень;
- параметри моделі можуть бути автоматично обрані таким чином, щоб описати наявний набір даних для навчання.

У системах розпізнавання мови приховані Марковські моделі зазвичай

застосовуються для подання фонем, або цілих слів. Кожний прихований стан представляє частину фонеми або слова. У кожен момент часу стан, в якому знаходиться система, може бути змінений відповідним набором перехідних ймовірностей, пов'язаних з даними станом [17].

2.3 Вибір інструментів та критеріїв для тестування методів розпізнавання мови

Для порівняння необхідно ввести деякі критерії, які змогли б відобразити важливі аспекти методів розпізнавання мови, такі як час виконання та кількість помилок.

В основі будь-якої мовної технології лежить так званий «engine», або ядро програми – набір даних та правил, за якими здійснюється обробка даних. Залежно від призначення цього ядра розрізняють TTS та ASR engine. TTS (Text-To-Speech) engine надає можливість синтезу мови за текстом, а ASR (Automatic Speech Recognition) engine – для розпізнавання мови. Існує кілька великих виробників, які займаються створенням ASR ядер.

Компанія SPIRIT займається створенням програмних засобів для цифрової телефонії, ущільнення мови, ідентифікації мовця для технологій VoIP та GPS. ASR engine від SPIRIT розроблений для розпізнавання мовних команд і застосовується в різних застосунках, таких як голосове управління пристроями, голосовий набір в hands-free пристроях, введення персональних ідентифікаційних кодів (PIN) в системах безпеки. Дане ядро вбудовується в будь-які DSP або RISC платформи і поставляється у вигляді об'єктного коду.

Sakrament ASR Engine – програмна розробка білоруської компанії «Сакрамент», розрахована на застосування в різних апаратних системах і програмних застосунках, що використовують технології розпізнавання мови. Заявлені характеристики:

- точність розпізнавання 95-98%;

- незалежність від диктора;

- незалежність від мови.

Однак, дана система не має можливості навчання – додаткові словники створюються за замовленням самою компанією «Сакрамент».

Sphinx – відкритий програмний продукт для розпізнавання мови. Розробка ведеться в університеті Карнегі-Меллона, продукт поширюється на умовах ліцензії Berkley Software Distribution (BSD) і доступний як для комерційного, так і для некомерційного використання. Його особливості:

- незалежність від диктора;

- розпізнавання безперервної мови.

Dragon NaturallySpeaking Preferred фірми Dragon Systems – єдина програма, яка наблизилася до того, щоб відповідати заявленим характеристикам. В цілому пакет дуже близько підходить до досягнення заявленої безпомилковості розпізнавання – 95%. Хоча пакет Dragon і поступається деяким з конкурентів у тому, що стосується переміщення по екрану, правки й форматування, він перевершує всіх у головному – здатності з першого разу правильно записувати вимовлені слова.

Отже саме цей програмний продукт буде використовуватися для дослідження зі застосуванням наступних моделей для певних методів.

Спеціалізовані модулі, які можна втілювати в різні текстові редактори:

- VoiceCode – дозволяє набирати чистий програмний код за допомогою голосових команд, не торкаючись клавіатури. VoiceCode дозволяє диктувати код природним чином, при цьому автоматично перетворює людську мову в специфічні програмні функції. Програма працює лише з однією мовою програмування Python, але її можна практично без проблем адаптувати під інші мови програмування .

- EmacsListen – програмний модуль, що виконує голосові функції текстового редактора GNU Emacs. Він постачається з граматикою ShortTalk, має підтримку розпізнавання й нормалізації тексту. Модуль можна використовувати для реалізації інших мовних інтерфейсів.

- Voice Grip – додатковий макрос для редактора Emacs, створений з метою

спрощення розпізнавання мови для програмістів.

- Java by voice – серія макросів для редактора Emacs, спроектовані для спрощеного введення коду мовою Java[23].

2.4 Створення експериментального стенду

Комп’ютерні програми у наш час можна віднести до однієї з двох категорій:

- клієнтські – у яких всі оперції виконуються на стороні ЕВМ на якій вони встановлені;
- клієнт-серверні – у яких всі операції виконуються на стороні сервера(сторонньої ЕВМ), а на стороні клієнта вони лише відображаються.

Відображення відбувається завдяки так званому графічному інтерфейсу користувача.

В залежності від задач, що стоять перед системою розпізнавання, доцільно використовувати ту, чи іншу технологію.

Клієнт-серверні системи підрозділяються на двох − і трирівневі. У першому випадку клієнт запитує сервіси безпосередньо у сервера, у другому запити обробляються проміжними серверами, які координують виконання клієнтських запитів з підлеглими їм серверами.

Взаємодія клієнтського і серверного процесів виконується за допомогою програмного забезпечення передачі даних. Воно складається з декількох рівнів програмного забезпечення, що дозволяють передавати дані і керуючу інформацію між клієнтами і серверами. Це програмне забезпечення зазвичай прив'язана до мережі. Всі клієнтські запити і відповіді сервера передаються по мережі в формі повідомлень, в яких містяться керуюча інформація і дані.

В даній роботі буде використовуватися клієнтський підхід.

Експериментальний стенд буде складатися з трьох частин:

- алгоритм DTW;
- алгорит побудований на прихованих Марківських системах;

- нейронні мережі.

Для точного визначення часу спрацьовування кожного з алгоритмів для них необхідно створити умови, при яких вони будуть нормально функціонувати. Кожний алгоритм потребує певних умов, при яких він покаже найбільший свій потенціал. Тобто для справедливого порівняння необхідно для кожного алгоритму налаштувати програму.

Для стандартного алгоритму динамічної трансформації часової шкали необхідно побудувати базу даних еталонів, для того щоб можна було справедливо оцінити його продуктивність.

Для алгоритму побудованому на прихованій Марківській системі необхідна кодова книга.

Для нейромережі необхідне її навчання. Без навчання, нажаль, нейромережа не працюватиме, що негативно скажеться на результатах експерименту.

2.5 Висновки до розділу 2

В даному розділі розглянуто та проаналізовано методи розпізнавання речі. Виділяючи переваги і недоліки кожного з методів можна прийти до висновку про кожний.

Виконано планування експеремнту та сформовані критері оцінення також визначелись з інстрементами проведення експеримету після цього було визначено палн обробки та інтерпретації отриманих даних. Для вирішення даного завдання будуть використані класичні методи аналізу та обробки: знайдені недоліки та переваги кожного з методів, та для оцінки розподілу значень метрик будуть побудовані гістограми.

3 АНАЛІЗ РЕЗУЛЬТАТІВ ЕКСПЕРИМЕНТАЛЬНОГО ДОСЛІДЖЕНЯ ТА ФОРМУВАННЯ ПРАКТИЧНИХ РЕКОМЕНДАЦІЙ ЩОДО ЕФЕКТИВНОСТІ МЕТОДІВ РОЗПІЗНАВАННЯ МОВИ

3.1 Рекомендації щодо удосконалення методів розпізнавання мови

3.1.1 Комбінований метод

Пропонується використовувати комбінований метод, який складається з трьох алгоритмів, що дозволить прискорити процес розпізнавання мови.

Для точного розуміння побудови комбінованого методу необхідно розглянути кожен з модифікованих для нього алгоритмів, які використовуються, більш детально. Це дозволить повністю зрозуміти принцип дії методу та за рахунок чого виконується прискорення розпізнавання звукових мовних сигналів.

Удосконалений Алгоритм DTW

В системах розпізнавання мовних сигналів, що основані на використанні алгоритму DTW, вхідний звуковий сигнал порівнюється з кожним еталоном почергово (рис 3.1).

Основним недоліком методу динамічної трансформації часової шкали є недостатня продуктивність для розпізнавання великих словників. При збільшенні кількості слів у словниці, тобто при занесенні нового слова в словник, кількість еталонних слів збільшується, що забезпечує необхідну точность.

На основі цього можна зробити висновок, що для модифікації необхідно зменшити кількості еталонів.

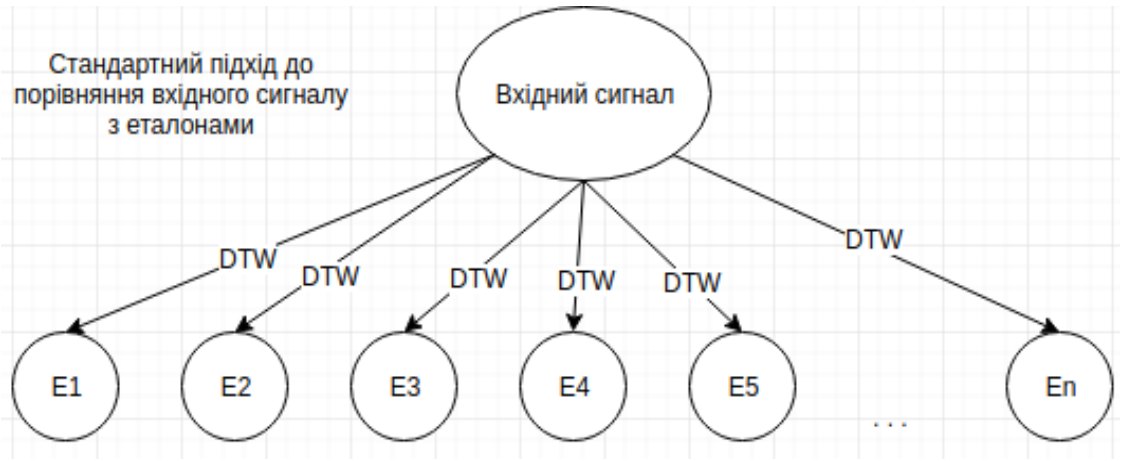

Рисунок 3.1 – Робота станартного алгоритму динамічної трансформації часової шкали.

Поєднання схожих між собою звукових сигналів у групи допоможе зменшити кількість еталонів та зменшити кількість порівнянь (рис. 3.2).

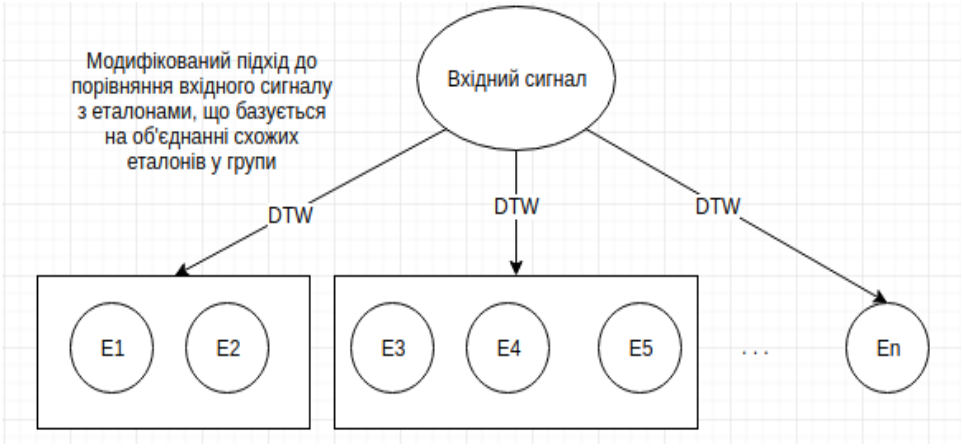

Рисунок 3.2 – Робота модифікованого алгоритму динамічної трансформації часової шкали.

Для того щоб модифікувати алгоритм необхідно розробити спеціальний блок для аналізу вхідних еталонів − аналізатор еталонів.

Одним з важливих недоліків цієї модифікації є необхідність проводити «навчання» системи. На сьогодняшній день даний алгоритм не може

гарантувати стабільну роботу будь-якого, вбудованого, словника, тому при використанні в комерційних системах, даний алгоритм необхідно тестувати для окремого випадку та модифікувати параметричні характеристики аналізатора вхідних звукових сигналів.

Об’єднуючи еталони в невеликі групи доцільно використовувати смислове значення мовного слова. Тобто всі еталонні записи, що містять в собі слово «two», будуть належати до однієї групи. У наш час досягти такого усереднення еталонів, щоб одна група еталонів представляла одне слово і як наслідок алгоритм DTW використовувався лише один раз (рис. 3.2) неможливо. У реальності модифікована система зберігання еталонів буде містити в собі декілька груп еталонів (рис. 3.3) і метод динамічної трансформації часової шкали необхідно буде застосовувати до визначення найкращого сходження.

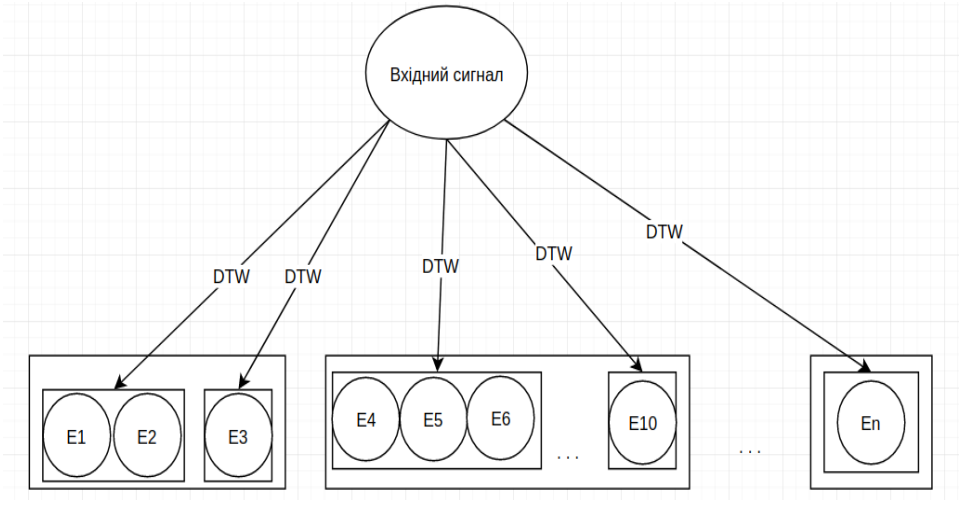

Рисунок 3.3 – Об’єднання еталонів для кожного слова у підгрупи.

З цього випливає, що у кожної підгрупи етолонів є представник який знаходиться шляхом усереднення всіх еталонів, що входять до підгрупи. В межах кожної підгрупи буде знаходитись лише один, еталон. Для простоти розуміння, надалі, якщо до групи входить один еталон, він все одно буде вважатися усередненим, хоча апроксимації здійснено не було. Усереднюється

новий, вхідний еталон з найкращим елементом підгрупи, якщо аналізатор вирішив додати його.

Для апроксимації середнього значення між звуковими сигналами буде використано метод найменших квадратів. Метод найменших квадратів − це математичний метод, що застосовується для вирішення різних завдань, заснований на мінімізації суми квадратів відхилень деяких функцій від шуканих змінних (рис. 3.4.).

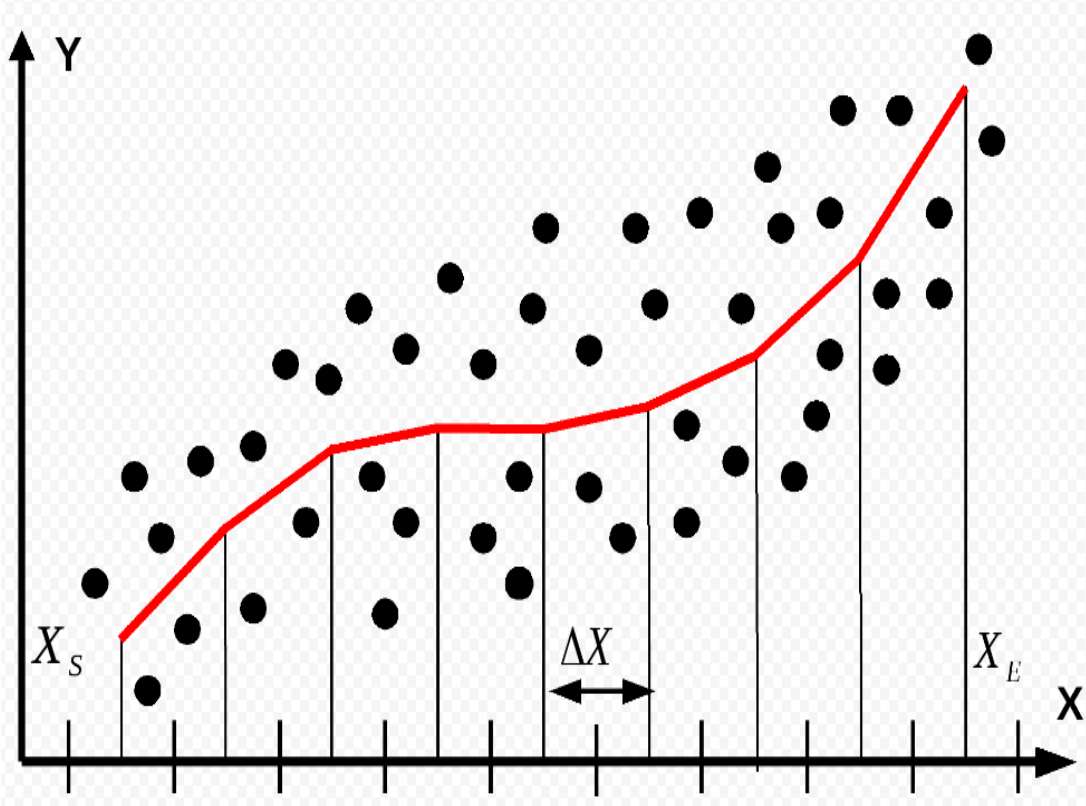

Рисунок 3.4 – Робота методу найменших квадратів.

Даний метод гарно підходить для апроксимації еталонів у підгрупах, але метод не враховує довжину часових рядів різних еталонів у підгрупах та групах.

Робота аналізатора еталонів. Перед тим як застосувати аналізатор еталонів, звуковий сигнал подається на систему параметризації (рис. 3.1), тобто набуває параметрів. Тоді на вхідний порт аналізатора еталонів надходить не надісланий

мовний сигнал, а очищений від шумів та придатний до аналізу.

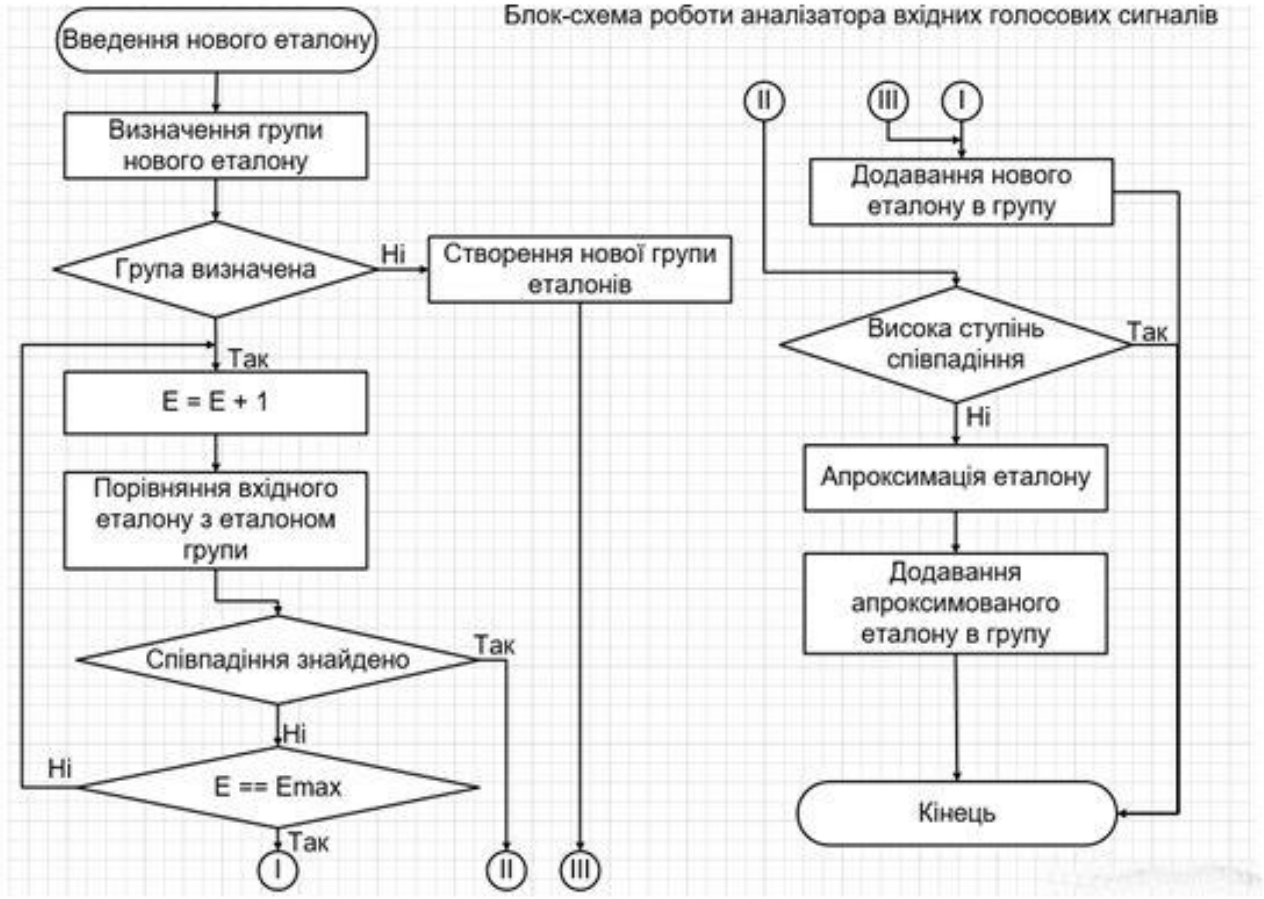

Рисунок 3.5 − Блок-схема роботи аналізатора вхідних голосових сигналів

Роботу аналізатора можна поділити декілька етапів (рис. 3.5):

- обробка сигналу методом динамічної трансформації часової шкали для встановлення подібності нового еталону до вже існуючих. Для вибору використовуються тільки ті підгрупи, до яких належить слово, до якого належить вхідний сигнал;
- при достатньо високому значенні коефіцієнта подібності вхідного сигналу хоча б з однією з підгруп, вхідний сигнал можна не включати до бази даних;
- при високому значення коефіцієнта подібності, вхідний сигнал можна усереднити з однією з підгруп еталонів;

- при малому значенні коефіцієнта подібності вхідний сигнал стає представником нової підгрупи.

Для роботи аналізатора необхідні константи, які будуть позначати для кожного етапу значення коефіцієнта подібності. Максимальним значенням констант − одиниця. Мінімальне значення − нуль.

Через те, що існує велика кількість різноманітних значущих параметрів та впливу шуму у вхідному сигналі, два вхідні еталони можуть сильно різнитися між собою. Тому значення константи, що відповідає за схожість доцільно обирати близьким до можливого значення.

Алгоритм побудований на закритій Марківській системі

Як вже будо сказано раніше, алгоритм побудований на закритій Марківській системі завжди видає результат, не залежно від того на скільки він правдивий. Тобто, якщо слово, яке потрібно розпізнати є в кодовій книзі з якою відбувається порівняння відсутнє, то алгоритм видає користувачу слово, яке найбільше на нього схоже. Для розробки комбінованого методу ця особливість не підходить, тому його треба модифікувати.

Модифікація алгоритму відбувається таким чином, закрита Марківська система замінюється на відкриту. Ця модифікація дозволяє добитися того, що в системі залишиться вихід для не розпізнаного слова. Завдяки цьому алгоритми описані вище можна буде скомбінувати.

Комбінований Метод

У запропонованому алгоритмі всі недоліки мінімізовані таким чином.

За основу методу взято алгоритм, заснований на Марківській системі, тому, що цей алгоритм є самим швидкодіючим із всіх вище перелічених. Але сама Марківська система перероблена таким чином, що при недосягненні певної ймовірності переходу слово або символ не є розпізнаними. Для мінімізації витрат часу і ресурсів використовується ступінчаста подача звукових сигналів. Під ступінчастою подачею мається на увазі поступове зменшення довжини подається на розпізнавання звукової хвилі, наприклад спочатку на розпізнання надходить слово, потім окремі його склади, потім

букви і якщо результат все одно не досягнуть в силу вступає алгоритм тимчасовий динамічний алгоритм (Рабінел - ламелі).

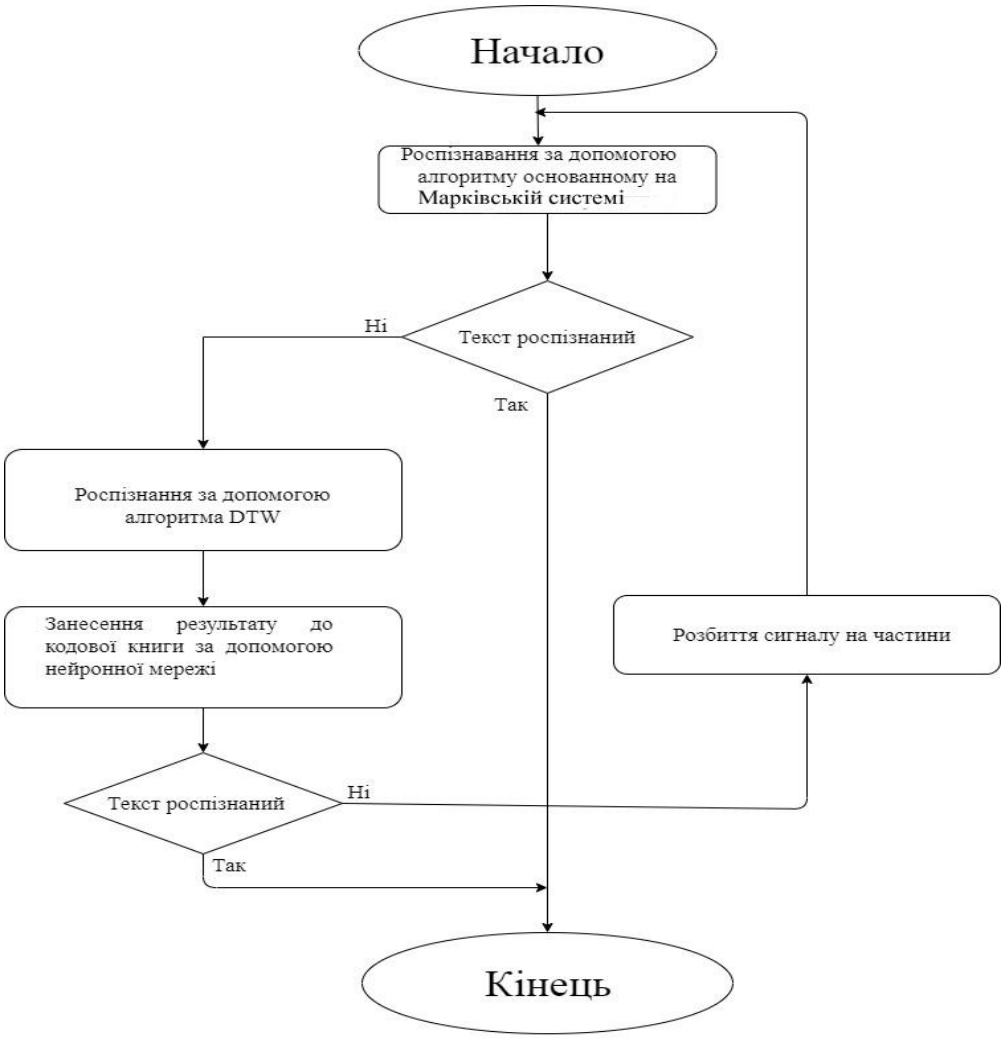

Рисунок 3.6 − Схема роботи комбінованого алгоритму

За рахунок такого підходу можна відсіяти зрозумілі для системи слова, зменшивши, тим самим кількість операція для їх розпізнання, але для слів, що не були розпізнані на етапі Марківського системи, шлях до розпізнавання значно збільшується. Для того, щоб уникнути дану проблему в комбінований алгоритм була впроваджена нейронна мережа. Кожен раз коли Марківська система не зможе визначити слово і для його визначення доведеться задіяти алгоритм DTW, нейромережа буде обробляти дане слово, зберігаючи його і

розділяючи на склади і літери і заносити в кодову книгу. Це рішення дозволить мінімізувати використання алгоритму DTW при тривалому застосуванні програми, а так само «привчити» програму до її активного користувача, так як інтонація, риси голосу, вимова у однієї і тієї ж людини в різний проміжок часу можуть відрізнятися один від одного [24].

3.2 Порівняння удосконаленого методу DTW з стандартним

3.2.1 Апробація результатів стандартного та модифікованого алгоритмів DTW

Необхідно визначити кількість еталонних слів, необхідних для задовільної точності розпізнавання використовуючи стандартний алгоритм DTW.

В таблиці 4.1, 4.2 та 4.3 наведені залежності точності та швидкості розпізнавання мовних сигналів від кількості еталонів для кожного слова. Кожне вхідне слово було проаналізовано 10 разів. Вхідне слово промовлялося з різним тембром, інтонацією та швидкістю.

Дивлячись на результати можна зрозуміти, що для простого слова, яке не має схожих за вимовою слів в словнику необхідно менше еталонів ніж для слова, яке має схожі по вимові слова.

Таблиця 4.1 − Результат слова «one» для станартного алгоритму динамічної трансформації часової шкали.

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	20мс	2
2	60мс	4
3	160мс	5
4	230мс	5
5	259мс	6

Продовження таблиці 4.1

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
6	300мс	6
7	320мс	7
8	360мс	8
9	800мс	9
10	1000 мс	10

Таблиця 4.2 − Результат слова «fury» для станартного алгоритму динамічної трансформації часової шкали.

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	22мс	2
2	70мс	4
3	179мс	5
4	237мс	5
5	391мс	6
6	443мс	6
7	534мс	7
8	652мс	8
9	801мс	9
10	1211 мс	10

Таке складне слово, як «fury» потребує велику кількість еталонів для точного розпізнавання, порівняно з коротким словом, але меншу за слово, яку має схожі слова. Чим більше схожих по хвилі слів буде в еталонному словнику тим більше слів потрібно буде зберігати для необхідної точності. Необхідно збільшувати кількість еталонних сигналів для тих слів, які мають схожі по вимові слова. Очевидно що при зменшенні кількості еталонів час на

розпізнавання зменшується пропорційно.

Далі визначимо кількість еталонів, необхідних для забезпечення точності розпізнавання використовуючи модифікацію алгоритму динамічної трансформації часової шкали. В таблиці 4.4, 4.5 та 4.6 наведено залежність точності та швидкості розпізнавання слів від кількості еталонів для кожного слова.

Таблиця 4.3 − Результат слова «data» для станартного алгоритму динамічної трансформації часової шкали.

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	21мс	2
2	63мс	4
3	171мс	5
4	241мс	5
5	388мс	6
6	456мс	6
7	587мс	7
8	722мс	8
9	833мс	9
10	1132 мс	10

Таблиця 4.4 − Результат слова «one» для модифікованого алгоритму динамічної трансформації часової шкали.

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	17мс	2
2	61мс	4
3	155мс	5

Продовження таблиці 4.4

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
4	212мс	5
5	350мс	6
6	467мс	6
7	552мс	7
8	678мс	8
9	789мс	9
10	980мс	10

Таблиця 4.5 − Результат слова «fury» для модифікованого алгоритму динамічної трансформації часової шкали..

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	21мс	2
2	62мс	4
3	164мс	5
4	233мс	5
5	398мс	6
6	459мс	6
7	555мс	7
8	690мс	8
9	823мс	9
10	1345 мс	10

Таблиця 4.6 − Результат слова «data» для модифікованого алгоритму динамічної трансформації часової шкали.

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	21мс	2
2	67мс	4
3	169мс	5
4	251мс	5
5	377мс	6
6	444мс	6
7	567мс	7
8	666мс	8
9	832мс	9
10	1119мс	10

Таблиця 4.7 − Результат слова «one» для стандартного алгоритму динамічної трансформації часової шкали.

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	19мс	2
2	55мс	4
3	175мс	5
4	260мс	5
5	313мс	6
6	400мс	6
7	540мс	7
8	621мс	8
9	769мс	9
10	1234мс	10

Таблиця 4.8 − Результат слова «fury» для стандартного алгоритму динамічної трансформації часової шкали.

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	19мс	2
2	69мс	4
3	148мс	5
4	257мс	5
5	386мс	6
6	498мс	6
7	500мс	7
8	721мс	8
9	821мс	9
10	964мс	10

Таблиця 4.9 − Результат слова «data» для стандартного алгоритму динамічної трансформації часової шкали..

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	20мс	2
2	60мс	4
3	160мс	5
4	230мс	5
5	360мс	6
6	451мс	6
7	556мс	7
8	781мс	8
9	800мс	9
10	1000 мс	10

Аналізуючи таблиці 4.7, 4.9, 4.1 та 4.3 видно, що для слів, які не мають схожих по вимові слів в словнику, модифіковані еталони забезпечують необхідну точність розпізнавання при зменшенні часу розпізнавання. Для слова, яке має схожі по вимові слова (таблиці 4.8 та 4.2) необхідна точність не забезпечена.

Необхідно провести розпізнавання слів використовуючи модифікований алгоритм динамічної трансформації часової шкали та модифіковані еталони. В таблиці 4.7, 4.8 та 4.9 наведено залежність швидкості та точності розпізнавання слів від кількості еталонів для кожного слова.

Таблиця 4.10 − Результат слова «one» для стандартного алгоритму динамічної трансформації часової шкали.

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	19мс	2
2	56мс	4
3	161мс	5
4	219мс	5
5	333мс	6
6	442мс	6
7	524мс	7
8	667мс	8
9	897мс	9
10	969 мс	10

Таблиця 4.11 − Результат слова «fury» для стандартного алгоритму динамічної трансформації часової шкали..

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	19мс	2
2	61мс	4
3	159мс	5
4	209мс	5
5	245мс	6
6	333мс	6
7	431мс	7
8	549мс	8
9	811мс	9
10	1187мс	10

Таблиця 4.12 − Результат слова «data» для стандартного алгоритму динамічної трансформації часової шкали.

Кількість еталонів для кожного слова	Швидкість аналізу одного сигналу, мс	Кількість правильних результатів
1	22мс	2
2	65мс	4
3	150мс	5
4	220мс	5
5	280мс	6
6	336мс	6
7	420мс	7
8	551мс	8
9	721мс	9
10	890 мс	10

Як можна помітити, при зменшенні кількості еталонів модифікований алгоритм DTW не забезпечує ту точність розпізнавання, яка необхідна при обробці слів, які мають схожі екземпляри. Для простих слів, які не мають схожих по вимові слів метод забезпечує необхідну точність розпізнавання.

3.3 Порівняльна характеристика методів розпізнавання мови

Порівняння параметрів розпізнавання мови на прикладі слів «Привіт» і «Мир» представлені в таблицях 4.13, 4.14.

Таблиця 4.13 − Результат по слову «Hello»

Метод	Середній час на розпізнавання слова, мс	Кількість правильно розпізнання слів зі 100 спроб
DTW	1400	80
Марковська система	980	63
Нейронна мережа	1200	75

Таблиця 4.14 − Результат по слову «World»

Метод	Середній час на розпізнавання слова, мс	Кількість правильно розпізнання слів зі 100 спроб
DTW	1130	83
Марковська система	830	65
Нейронна мережа	940	76

Як видно з результатів експерименту продуктивність алгоритму динамічної трансформації часової шкали вище не на багато.

Таблиця 4.15 − Результат по реченню «I am very hungry, please can somebody feed me»

Метод	Середній час на розпізнавання речення, мс	Кількість правильно розпізних речень зі 100 спроб
DTW	2130	79
Марковська система	1430	58
Нейронна мережа	2360	75

Як можна помітити алгоритм динамічної трансформації часової шкали не втрачає своє лідерство і на етапі розпізнавання речення.

3.4 Висновки до розділу 3

В даному розділі проаналізовано результати експериментів з методів для розпізнавання мови і було виявлений найкращий після цього були сформовані рекомендації щодо покращення даних методів та проведено апробацію за модифікованими методдами.

Також були сформовані рекомендації по створенню комбінованого методу розпізнавання мови, який дозволяє прискорити процес трансформації голосових сигналів у текстові дані.

ВИСНОВКИ

В магістерській роботі виконано експериментальне дослідження, був спланований експеримент та проаналізовані його результати з яких було сформульовано певні рекомендації щодо удосконалення методів розпізнавання мови а також запропоновано комбінований метод. А також в даній роботи було проведено аналіз існуючих алгоритмів для розпізнавання мови. В роботі були розглянуті такі алгоритми як :

- тимчасові динамічні алгоритми (Dynamic Time Warping). Контекстно-залежна класифікація. При її реалізації з потоку мови виділяються окремі лексичні елементи − фонеми і Алофон, які потім об'єднуються в склади і морфеми;
- приховані Марківські моделі (Hidden Markov Model);
- нейронні мережі (Neural networks).

На основі проведеного аналізу було зроблено висновок про те що дані алгоритми дозволяють розпізнавати мову, але використовуючи їх окремо один від одного продуктивніть значно знижується.

Було запропоновано комбінований алгоритм розпізнавання слів. Головною перевагою цього методу є прискорення розпізнавання слів, завдяки правильному комбінуванню вже існуючих алгоритмів та значному модифікування одного з них.

З не достатків цього методу хотілося б виділити значно більшу кількість часу на його налаштування та доведення його до стану працездатності, адже він містить у собі три алгоритми, які потребують індивідуального підходу

ПЕРЕЛІК ПОСИЛАНЬ

1. Вишнякова О. А., Лавров Д. Н. Применение преобразования Гильберта-хуанга к задаче сегментации речи [Текст] // Математические структуры и моделирование. 2011. вып. 24. С. 12–18.

2. Davies, K. H., Biddulph, R. and Balashek, S. (1952) Automatic Speech [Текст] Recognition of Spoken Digits, J. Acoust. Soc. Am. 24 (6) 637 – 642 с.

3. Винцюк Т. К. – Анализ, распознавание и интерпретация речевых сигналов. [Текст] – Киев 1985 г.

4. Рабинер Л. Р., Шафер Р. В. – Цифровая обработка речевых сигналов – [Текст] - М.: Радио и связь, 1981 г.

5. Современные проблемы в области распознавания речи. - Auditech.Ltd. [Текст] Бысько М. В. – Шумология – "Медиамузыка". – 2014. – № 3 Тэйлор Р. – Шум. – М.: Мир, 1978.

6. Чекмарьов А. Мовні технології - проблеми та перспективи. [Текст] // Компьютерра, № 49 с. 26-43, 1997 р.

7. Фролов А., Фролов Г. Синтез и распознавание речи. [Текст] Современные решения. – 2003.

8. Барабаш Ю. Л., Зінов'єв Б. В Питання статичної теорії розпізнавання. - М .: [Текст] Сов. радіо, 1967.- 400 с.

9. Churyumov Genadiy Method for Ensuring Survivability of Flying Ad-hoc Network Based on Structural and Functional Reconfiguration / Genadiy Churyumov, Vitalii Tkachov, Volodymyr Tokariev, Vladyslav Diachenko [Текст] // Selected Papers of the XVIII International Scientific and Practical Conference “Information Technologies and Security” (ITS 2018) / Kyiv, Ukraine, November 27, 2018. – Pp. 64-76.

10. Dynamic Programming Algorithms in Speech Recognition – Titus Felix – FURTUNĂ Academy of Economic Studies, [Текст] Bucharest.

11. Математические методы распознавания образов – МГУ, ВМиК, кафедра «Математические методы прогнозирования» – [Текст] Местецкий

Леонид Моисеевич, 2002–2004.

12. Volodymyr Tokariev Ultra Wideband Signals in Control Systems of Unmanned Aerial Vehicles / Aleksandr Serkov, Valeri Kravets, Igor Yakovenko, Gennady Churyumov, Wang Nannan [Текст] // The 10h IEEE International Conference on Dependable Systems, Services and Technologies, DESSERT'2019 5-7 June, 2019, Leeds, United Kingdom. – Pp.26 – 29. ,

13. Vaswani N. Principal components null space analysis for image and video classification / N. Vaswani, R. Chellappa [Текст] // IEEE Trans. Image Process. – 2006. – Vol. 15, No. 7. – P. 1816–1830.

14. Suhas S. Face recognition using principal component analysis and linear discriminant analysis on holistic approach in facial images database / S. Suhas, A. Kurhe, Dr.P. Khanale[Текст] // IOSR Journal of Engineering. – 2012. – Vol. 2, Is. 12. – P. 15-23.

15. Bayesian Approaches in Speech Recognition – Shinji Watanabe – [Текст] NTT Communication Science Laboratories, NTT Coporation, Kyoto, Japan.

16. Ghazi Al-Naymat, Sanjay Chawla, Javid Taheri Sparse DTW: [Текст] A novel approach to speed up Dynamic Time Warpingю

17. К.В. Сидоров, Н.Н. Филатова – АНАЛИЗ ПРИЗНАКОВ ЭМОЦИОНАЛЬНО ОКРАШЕННОЙ РЕЧИ – [Текст] Вестник ТвГТУ, 180 (Вып. 20). стр. 26-32. ISSN 2224-6363.

18. Честович Л. А., Венцов А. В., и др. [Текст] – Восприятие речи человеком – Академия наук СССР.

19. Klass, Philip J. Fiber Optic Device Recognizes Signals. // Aviation Week & Space Technology. – N.Y.: [Текст] McGraw-Hill, 1962. – Vol. 77 – No. 20 – P. 94-101.

20. Eamonn J. Keogh, Michael J. Pazzani [Текст] – Derivative Dynamic Time Warping, Section 1.

21. Stan Salvador and Philip Chan – Fast DTW: [Текст] Toward Accurate Dynamic Time Warping in Linear Time and Space.

22. Norbert Wiener – Cybernetics or Control and Communication in the

Animal and the Machine. – [Текст] (Hermann & Cie Editeurs, Paris, The Technology Press, Cambridge, Mass., John Wiley & Sons Inc., New York, 1948)

23. Агарков М.О., Ільїна І.В., Сумцов Д.В. [Текст] –«Застосування технологій розпізнавання мови в розробці програмного забезпечення» – Наукове видання ПРОБЛЕМИ ІНФОРМАТИЗАЦІЇ – 2019, ст. – 6.

24. Belhumeur P.N. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection / P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman [Текст] // IEEE Trans. On PAMI. – 1997. – Vol. 19, No. 7. – P. 711–720.

25. Yang M.H. Kernel Eigenfaces vs. Kernel Fisherfaces: face recognition using kernel methods / M.H. Yang [Текст] // 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition. – 2002. – P. 215-220.

26. Bartlett M.S. Face recognition by independent component analysis / M.S. Bartlett, J.R. Movellan, T.J. Sejnowski [Текст] // IEEE Trans. Neural Netw. – 2002. – Vol.13, No. 6. – P. 1450–1464.

27. Shen L. A review on Gabor wavelets for face recognition / L. Shen, L. Bai [Текст] // Journal of Pattern Analysis and Applications. – 2006. – Vol. 9, No. 2-3. – P. 273- 292.

28. Imtiaz H. A face recognition scheme using waveletbased dominant features / H. Imtiaz, S.A. Fattah [Текст] // Signal & Image Processing : An International Journal. – 2011. – Vol.2, No.3. – P. 69-80.

29. Zhao W. Face recognitions literature survey / W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld [Текст] // ACM Computing Surveys. – 2003. – Vol. 35, No. 4. – P. 399–458.