1 АНАЛІЗ ЗАДАЧІ РОЗПІЗНАВАННЯ МОВНИХ КОМАНД В СИСТЕМАХ ЛЮДИНО-МАШИННОЇ ВЗАЄМОДІЇ

1.1 Опис загальних положень

Мова — це одна з найбільш природних способів передачі інформації між суб'єктами мовлення.

Мовні команди — це слова однієї з природних або штучних мов, які представляються аудіо сигналом та використовуються у людино-машинній взаємодії. Кожна команда може складатися з одного або декілька слів

Мовлення було природним елементом комунікації між людьми ще багато тисяч років тому. Завдяки великій кількості слів, що складають мову, цей механізм взаємодії є незамінним при забезпеченні передачі великої кількості інформації та є одним з найбільш швидких з усіх можливих природних інтерфейсів взаємодії.

ЛМВ (людино-машинна взаємодія) — це вивчення, планування та розробка взаємодії між людьми (користувачами) і комп’ютерами. Найчастіше її розглядають як сукупність науки про комп’ютери, науки про поведінку, проектування та інших областей дослідження. Взаємодія між операторами і комп’ютерами відбувається на рівні інтерфейсу користувача, який включає в себе програмне та апаратне забезпечення; наприклад, образи чи об’єкти, що зображуються на екранах дисплеїв, дані, отримані від користувача за допомогою апаратних пристроїв введення (таких як клавіатури та миші) та інші взаємодії користувача з комп’ютерними системами [6].

Людино-машинна взаємодія займається [6]:

- методологією і розвитком проектування інтерфейсів (тобто, виходячи з вимог і класу користувачів, проектування найкращого інтерфейсу в заданих рамках, оптимізація під необхідні властивості, такі як здатність до навчання і ефективність використання);

- методами реалізації інтерфейсів (наприклад, програмні інструментарії, бібліотеки та раціональні алгоритми);
- методами для оцінки та порівняння таких інтерфейсів;
- розробкою нових інтерфейсів і технологій взаємодії;
- розвитком описових і прогнозованих моделей, і теорією взаємодії.

Сьогодні, для взаємодії людини та комп'ютера використовуються, переважно, такі пристрої, як миша та клавіатура, а також сенсорна панель. Комп'ютерна миша та клавіатура використовувались, як засоби взаємодії з ЕОМ (електронною обчислювальною машиною) ще за часів виникнення перших механізмів ЛМВ. Хоча їх зовнішній вигляд та легкість використання значно змінилися, але сутність їх залишилася тією ж самою. Тому можна сказати, що ці методи взаємодії з комп'ютером є досить застарілими.

На сьогоднішній день створюється велика кількість різноманітних систем ЛМВ, частиною яких є системи керування технічними засобами. Більшість з них направлена на покращення існуючих систем взаємодії для полегшення роботи користувача.

Мовна взаємодія має потенціал замінити собою усі інші способи комунікації з технічними системами за рахунок більшої швидкості розпізнавання однієї команди у порівнянні з усіма іншими наявними методами взаємодії, окрім нейроінтерфейсів, що отримують інформацію безпосередньо від мозкової активності людини. Однак усі переваги нейронних інтерфейсів не переважують значних недоліків, серед яких важкість у створенні подібних систем та необхідність у спеціальному обладнанні для роботи з мозковою активністю.

Зараз, в період загальної комп'ютеризації та глобальної інформатизації, на перше місце виходить проблема створення такого інтерфейсу користувача, який не тільки б задовольняв усі вимоги до ефективної взаємодії людини з комп'ютером, а й значно б її полегшував та покращував результативність завдяки якнайбільшій відповідності програмного забезпечення природній

взаємодії людини та його відповідності всім особливим потребам користувача [7]. Розробка таких систем ЛМВ займає досить велику частину від усіх досліджень відомих компаній, зокрема найвідоміші з них є Google, Yandex та Microsoft.

Керування технічними пристроями за допомогою мовних команд є досить зручним та навіть є майже єдиною можливістю людино-машинної взаємодії для людей з певними вадами, що не можуть використовувати звичайні засоби взаємодії. Це, наприклад, люди з порушенням зорової активності або опорно-рухової системи, а також такі системи є корисним тренуванням для людей з певними порушеннями мовної активності.

Більшість сигналів (мовних в тому числі) мають аналогову природу, тому для обробки їх на цифрових комп'ютерах вони перетворюються в дискретні сигнали за допомогою АЦП (аналого-цифрового перетворювача). За допомогою цієї процедури отримують набір відліків s[n], знятих в моменти Δt*n миттєвих значень безперервного сигналу, які вже позбавлені фізичної природи, а їх максимальне і мінімальне значення задається розрядністю АЦП. Наприклад, якщо
розрядність АЦП дорівнює 2 байтам, то все значення у відліках укладаються в проміжок $[2^{16-1}, 2^{16-1}-1]$. При цьому найважливішим параметром перетворення є частота дискретизації, яка визначає скільки миттєвих значень безперервного сигналу (відліків) буде збережено за одну секунду.

Для опису та перетворення дискретних сигналів застосовуються засоби цифрової обробки сигналів (ЦОС). Найважливішою процедурою ЦОС є дискретне перетворення Фур'є (ДПФ) [8]:

$$S[m] = \sum s[n] * e^{\frac{j2\Pi nm}{N}}, m = 1,..,N \qquad (1.1)$$

де N — кількість відліків, за якими будується ДПФ;

j — уявна одиниця.

ДПФ дозволяє перейти з часової області в частотну, тобто розкласти s[n] на набір гармонік і знайти залежність амплітуди (енергії) гармоніки від її частоти. На рис. 1.1 представлена частина мовного сигналу з командою «видалення»:

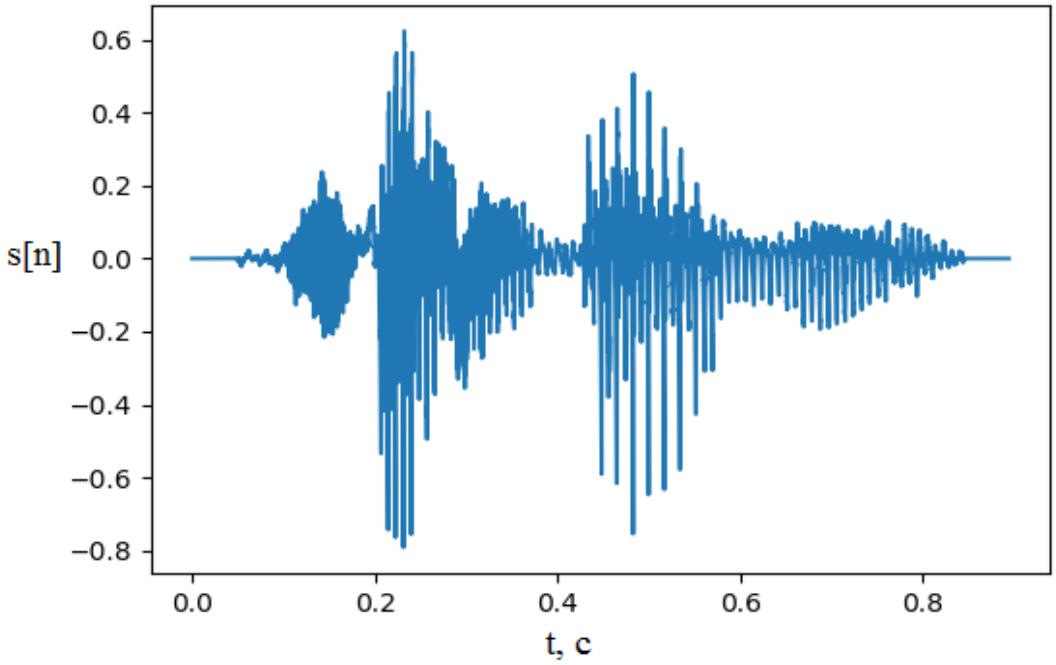

Рисунок 1.1 — Мовний сигнал з командою «видалення»

При цьому для того, щоб абстрагуватися від розрядності АЦП, відліки оцифрованого сигналу прийнято зображати у відносних величинах: або в частках від максимального значення, або в децибелах. ДПФ мовного сигналу з командою «видалення» зображене на рис. 1.2.

Для знаходження ДПФ була виділена ділянка розміром N = 1024 відліку. При цьому по горизонталі відкладається частота гармонік, а за вертикаллю — |S[n]|, що є амплітудою гармоніки.

Мова є нестаціонарним сигналом, тобто її характеристики змінюються з часом. Можна наочно зобразити ці зміни, побудувавши графіки модулів ДПФ для фрагментів (фреймів) мовного сигналу, що розташовані підряд.

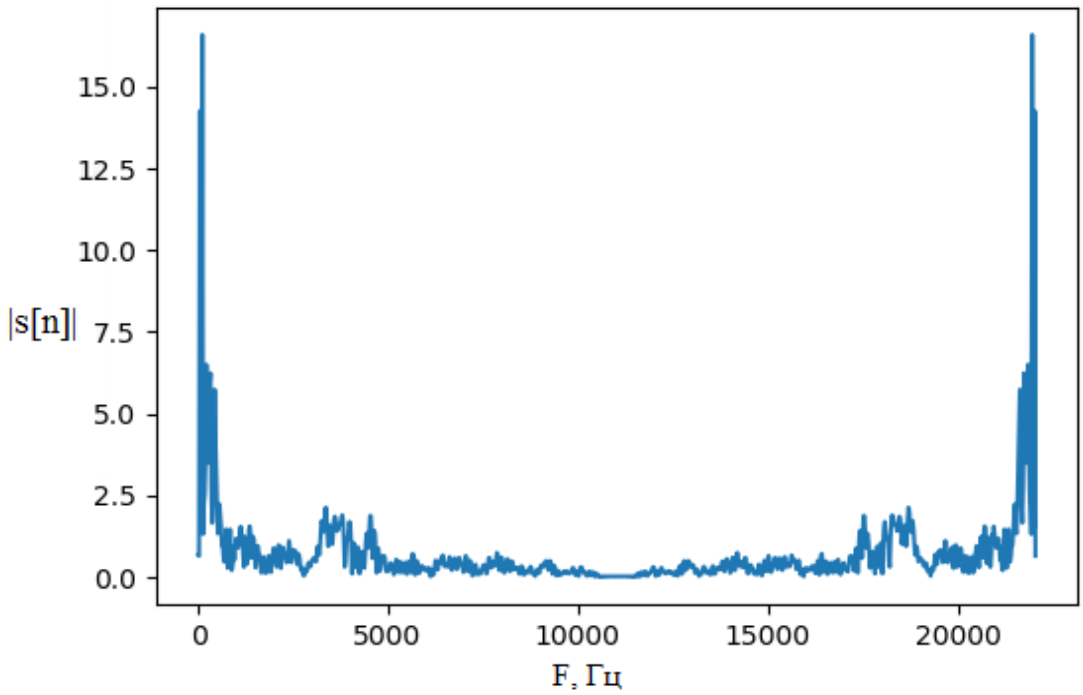

Рисунок 1.2 — ДПФ для частини мовного сигналу з командою «видалення»

Зображення модулів ДПФ називається спектрограмою (рис. 1.3).

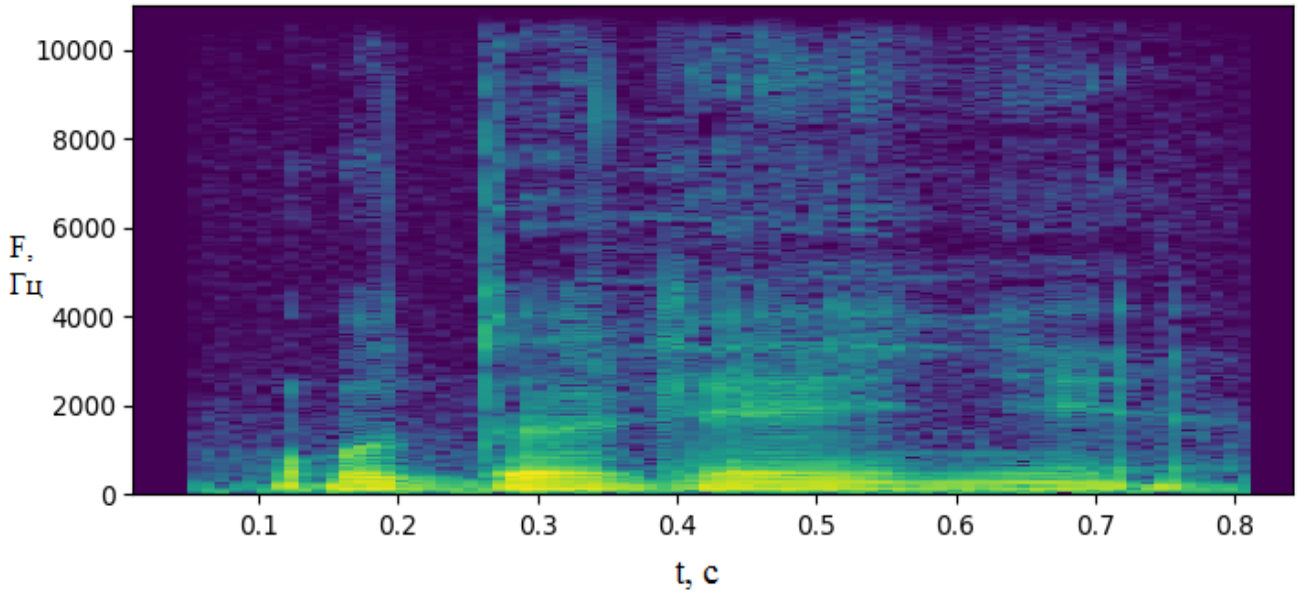

Рисунок 1.3 — Спектрограма частини мовного сигналу з командою «видалення»

По горизонтальній осі спектрограми відкладається час, по вертикальній - частота, а амплітуда відображається яскравістю або кольором. Спектрограми широко використовуються для аналізу звукових сигналів та музики, обробки мовних сигналів та розпізнаванні мови.

1.2 Аналіз існуючих систем ЛМВ з використанням мовних команд

У більшості систем розпізнавання мови використовуються окремі склади або літери слів у якості морфем, у тому числі в [2], у основі якого використання прихованих марковських моделей. Цей підхід має низку переваг, наприклад: досить малий обсяг бази даних для зберігання аудіо представлення морфем. Проте недоліком його є велика ймовірність неправильного розпізнавання при спотворенні частини сигналу перешкодою через малу тривалість аудіо фрагмента морфеми. Іншим недоліком є те, що наступна можлива морфема обирається після визначення попередньої, тобто достовірність визначення наступного фрагменту залежить від правильності розпізнавання попереднього. Неправильне розпізнавання команди може призвести до невиправних наслідків, таких як втрата важливих документів.

Такі системи створюються для визначення великою кількості слів звичайної мови людини. Зазвичай це системи, що перетворюють мовлення у текстове представлення для подальшого використання. Їх навчання відбувається за допомогою великої кількості аудіо фрагментів одної мови або зібраних заздалегідь її статистичних особливостей [2], тобто можливих комбінації морфем та їх положень у слові.

Поширеним є також розробка подібних систем з використанням Google Voice або інших сервісів автоматичного розпізнавання []. Такий підхід не можна використовувати у системі ЛМВ через неможливість використання її у місцях, де немає підключення до мережі Інтернет або є перебої з доступом до неї. Це також може стати причиною неправильного розпізнавання команд при

перебою. До того ж усе розпізнавання відбувається за допомогою цього сервісу автоматично, тобто неможлива зміна внутрішніх параметрів системи, способів обробки вхідних аудіо даних тощо. Іншим значним недоліком є необхідність покупки можливості використання деяких з сервісів для забезпечення розпізнавання.

Штучні нейронні мережі (ШНМ) використовуються у подібних задачах дуже часто. Цей підхід має низку переваг, серед яких висока швидкість розпізнавання та можливість виконувати подальше донавчання мережі, але він має й деякі значні недоліки. Навчання мережі займає багато часу в залежності від об’єму вхідних даних та обраного алгоритму навчання, а також від обраних параметрів мережі. Якість розпізнавання великою мірою залежить від вхідних даних, за допомогою яких відбувається навчання. Для забезпечення більшої достовірності розпізнавання команд необхідно проводити навчання на великій вибірці, але знаходження даних для цієї вибірки може бути дуже важкою справою в залежності від складності задачі [].

Таким чином, можна виділити такі системні недоліки: недостатня ефективність роботи існуючих систем, яка виражається у великій кількості неправильно визначених команд, а також велика ресурсоємність властива деяким з систем розпізнавання мовних команд, що виражається у необхідності використання спеціальних апаратних засобів, наприклад мікрофону високої якості, або у необхідності сплатити за використання цієї системи, що робить неможливим використання системи для деяких людей.

1.3 Аналіз методів отримання ознак з аудіосигналу

Розпізнавання мовних команд потребує аналізу збіжності кожної з команд до деяких еталонних значень з метою виділення найбільш схожої на вхідну команду.

Для цього необхідно отримання деяких даних з аудіосигналу команди, які однозначно характеризують команду та дозволяють відрізнити одну команду від другої. Використання необроблених даних аудіосигналу потребує аналіз великої кількості даних, при цьому розпізнавання не буде відповідати необхідним потребам до достовірності розпізнавання. Тому з аудіо даних виділяють характерні ознаки, проводити маніпуляції з якими набагато легше. Це також збільшує швидкість обробки та розпізнавання мовних команд досить сильно. Існує сотні характерних ознак аудіо сигналу та сотні способів отримання цих ознак. Тому був проведений порівняльний аналіз основних методів отримання ознак сигналу, що представлений у табл. 1.1.

Таблиця 1.1 — Порівняльний аналіз методів отримання ознак аудіо сигналу

Номер методу	Назва методу	Переваги	Недоліки
1	Лінійне прогностичне кодування	Нормальна роботи при низькій швидкості передачі звуку.	Велика варіативність результату на кожному наступному кадрі мови.
2	Мел-кепстральні коефіцієнти	Є найбільш поширеним та найчастіше використовуваним.	Велика чутливість до шуму на сигналі.
3	Швидке перетворення Фур’є	Забезпечує більшу, у порівнянні з іншими методами, інформаційність.	Ця інформація знаходиться на значно більшій частоті, ніж використовується звичайно, тож її не часто використовують.
4	Комбінування лінійного прогностичного кодування та мел-кепстрального коефіцієнту	Забезпечує більшу достовірність визначення ознак.	Незначно довше знаходження ознак у порівнянні з іншими методами.

Виходячи з даних у табл. 1.1 було обрано комбінування методів лінійного прогностичного кодування та мел-кепстрального коефіцієнту для використання при отриманні ознак, через те, що він забезпечує більшу надійність отримання ознак, а отже і достовірність розпізнавання команд.

1.4 Аналіз методів розпізнавання команд

Можна виділити три основних групи методів розпізнавання мови:

- методи, засновані на порівнянні з еталоном;
- методи, які виконують побудову вирішальних функцій;
- приховані марковські моделі.

1.2.3.1 Методи, засновані на порівнянні з еталоном

Для кожного слова складається модель-еталон промовляння O', щоб на етапі розпізнавання вибрати ту модель, еталон якої найближче до розглянутої акустичної послідовності O.

Головна проблема методів цієї групи полягає в тому, що мовні образи сильно відрізняються за тривалості, отже необхідний спосіб порівнювати образи різної довжини. Для цього використовується метод динамічного вирівнювання часу (Dynamic Time Warping — DTW) [10]. У ньому проблема різниці довжини еталона і розглянутого способу вирішується таким шляхом: складається матриця C розміром M × N, де N — довжина зразка, а M — довжина даної послідовності та виконується обчислення елементів цієї матриці:

$$\begin{gathered} C_{1,1} = D_{1,1} \\ C_{i,1} = D_{i,1} + C_{i-1,1},\ i = 2, ..., M \\ C_{1,j} = D_{1,j} + C_{1,j-1},\ j = 2, ..., N \\ C_{i,j} = D_{i,j} + \min(C_{i-1,j}, C_{i-1,j-1}, C_{i,j-1}),\ i = 2, ..., M;\ j = 2, ..., N, \end{gathered} \tag{1.2}$$

де $D_{i,j}$ — це відстань між i-м компонентом O' та j-м компонентом O, яке може обчислюватися різними способами, наприклад, як евклідова (1.3) або манхеттенська відстань (1.4):

$$D_{i,j} = \sqrt{o_i^2 + o_j^2} \tag{1.3}$$

$$D_{i,j} = \left| o_i - o_j \right| \tag{1.4}$$

Розглянутий метод, фактично, являє собою рішення задачі пошуку найкоротшого шляху на графі методом динамічного програмування, де початковий вузол розташований в лівому нижньому кутку сітки, а кінцевий - у правому верхньому з координатами (o'_N, o_M).

Недоліком методу динамічного вирівнювання часу є труднощі, що виникають при складанні еталона, які викликані сильною варіативністю мови. Але на сьогоднішній день розроблено різноманітні способи покращення роботи цього методу, що значно зменшують його недоліки у порівнянні з іншими представленими методами.

1.2.3.2 Методи, які виконують побудову вирішальних функцій

Сутність методів даної групи полягає в знаходженні такої функції, яка б за вхідним образом визначала його приналежність до того чи іншого класу. Для цього найбільш часто використовуються штучні нейронні мережі (Artificial Neural Networks - ANN). Одношаровий персептрон дозволяє побудувати розділяючи площині для лінійно-роздільних класів. Мінімальна обчислювальна одиниця персептрона — j-й штучний нейрон, який визначається як лінійна функція (1.5) з вагами $w_j = (w_{0j}, w_{1j}, ..., w_{Nj})$ від N аргументів, на які подається мовний образ $O = (o_1, o_2, ..., o_N)$ [10]:

$$y_j = w_{0j} + \sum_{i=1}^{N} w_{ij} o_i = w_j x^T = g_j(x), \quad (1.5)$$

де $x = (1, o_1, o_2, ..., o_N)$.

Кожен вихідний нейрон зіставляється одному з s класів ($\omega_1, \omega_2, ..., \omega_s$). У матричному вигляді це можна записати за допомогою матриці вагових коефіцієнтів персептрона $W = (w_1^T, w_2^T, ..., w_s^T)$ та вектору вихідних значень $y=(y1, y2, ..., y_s)^T$:

$$y = g(x) = W^T * x \quad (1.6)$$

Тоді для розпізнавання необхідно буде знайти той нейрон k, чиє вихідне значення буде максимальним:

$$k = ArgMax_{1<i<s}\, g_i(x) \quad (1.7)$$

В процесі навчання персептрона на навчальній вибірці налаштовують матрицю вагових коефіцієнтів таким чином, щоб мінімізувати помилку його відповіді.

Багатошаровий персептрон включає в себе один і більше прихованих шарів і дозволяє будувати нелінійні розділяючи функції. Для його навчання можна використовувати алгоритм зворотного поширення помилки [11].

Широкий інтерес до нейронних мереж викликаний їх здатністю до виділенню характерних рис способу і узагальненню [11]. Також плюсом можна вважати те, що штучний нейрон досить просто реалізувати апаратно, і, з'єднавши нейрони в мережу потрібної конфігурації, можна побудувати нейрокомп'ютер.

1.2.3.3 Приховані марковські моделі

За допомогою ланцюга станів ПММ моделюють фонеми мови, які, в свою чергу, об'єднують в слова. Найбільш адекватною вважається модель фонеми з трьох станів: початкового, середнього та кінцевого. Також зазвичай виділяють окремий стан під тишу і неінформативні звуки, наприклад, вдихи і видихи. При цьому вихідні ймовірності моделюються за допомогою моделей гауссовових сумішей (Gaussian Mixture Models - GMM).

Складають або окремі ПММ для кожного слова розпізнається словника, або одну велику ПММ, що об'єднує слова в реченні і більші структури. У першому випадку розпізнавання можна виконати за допомогою алгоритму прямого ходу (forward algorithm), знайшовши таку ПММ, яка здатна породити розглянуту послідовність з найбільшою ймовірністю [10]. У другому випадку, використовуючи алгоритм Вітербі, знаходять найбільш ймовірний ланцюг станів, через які повинна пройти ПММ для породження даної послідовності.

Другий підхід використовується частіше, так як з його допомогою можна розпізнавати злиту мова. Перевагою ПММ перед іншими методами є природне вбудовування часу в модель λ, що дозволяє врахувати варіативність промовляння по довжині і швидкості, а також перейти до розпізнавання злитої мови. Крім того, розроблені ефективні алгоритми ПММ, за допомогою яких можна здійснити розпаралелювання. Але, як було зазначено раніше, цей метод не є придатним для реалізації поставленої мети.

1.5 Змістова постановка задачі розпізнавання мовних команд у системах людино-машинної взаємодії

Напрямком даного дослідження є вивчення застосування різних методик, підходів та інформаційних технологій для обробки та розпізнавання мовних команд з аудіосигналу.

Метою роботи є розробка та дослідження методики розпізнавання мовних команд для збільшення достовірності розпізнавання в системах людино-машинної взаємодії.

Об’єктом дослідження є процес розпізнавання мовних команд в системах людино-машинній взаємодії.

Предметом дослідження є методика розпізнавання мовних команд в системах людино-машинної взаємодії.

Для досягнення поставленої мети в роботі необхідно вирішити наступні задачі:

- проаналізувати сучасні методи та засоби проектування систем ЛМВ та підходів до розпізнавання мовних команд, а також роботу існуючих системи ЛМВ, які використовують ці підходи;
- розробити методику розпізнавання мовних команд для покращення розпізнавання команди з аудіосигналу;
- розробити програму для реалізації попередньо створеної методики розпізнавання мовних команд та перевірки ефективності її функціонування у порівнянні з іншими підходами до розпізнавання мовних команд;
- провести експериментальне дослідження розробленої програми, яка реалізує роботу методики та проаналізувати достовірність розпізнавання мовних команд з її використанням.

1.6 Висновки до першого розділу

Аналіз, проведений у першому розділі, показав, що використання прихованих марковських моделей для системи розпізнавання роздільних

команд не дасть потрібного рівня достовірності розпізнавання, адже для нього важливим є контекст, тобто попередньо розпізнані слова або фонеми. У системі з використанням роздільних від контексту команд дуже важко буде розрахувати статистичну імовірність наступної фонеми, що вкрай негативно вплине на достовірність розпізнавання.

Використання штучних нейронних мереж потребує великий об'єм бази даних для навчання системи та виконання адекватного розпізнавання команд. Збір великої кількості даних не завжди є можливим на початку побудови систем ЛМВ. Приймаючи до уваги ще час, необхідний для навчання мережі, можна зробити висновок про те, що подібний підхід не є кращім для поставленої мети.

Найбільш придатним для використання у інформаційній системі ЛМВ на базі мовних команд є метод порівняння з шаблонами. Такий підхід дозволяє зберегти достатню для розроблюваного інтерфейсу ЛМВ достовірність розпізнавання команд за рахунок більшої тривалості сигналу команди, ніж при використанні фонем.

Одним з недоліків цього методу є важкість побудови розпізнавання злитою мови, тобто такої, коли кожна команда промовляється без пауз після попередньої. Проте, цей недолік не є значним при побудові даної системи, через те, що в ній будуть розпізнаватися лише роздільно вимовлені команди. Це зроблено для забезпечення найбільшої достовірності розпізнавання кожної команди.

Мова людини є дуже варіативною навіть для однієї і тієї ж людини, що значно впливає на достовірність результату розпізнавання. Тому необхідне використання алгоритмів, що дозволяють оцінити збіжність різних за розміром, швидкістю промовляння та іншими параметрами аудіо даних.

Таким чином, необхідно створити методику віднесення кожної команди за її отриманим аудіосигналом та виконаної обробки до певної групи еталонів команд.