ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ФРАНКА
Факультет прикладної математики та інформатики
Кафедра дискретного аналізу та інтелектуальних систем

ДИПЛОМНА РОБОТА

на тему:

РОЗПІЗНАВАННЯ ПРИРОДНОЇ МОВИ ШТУЧНИМ ІНТЕЛЕКТОМ

Студента IV курсу, групи ПМі-45,
спеціальності 122 — комп'ютерні науки
Зубальського А. М.
Керівник професор Щербина Ю. М.

Національна шкала _____
Кількість балів: _____
Оцінка: ECTS _____

Львів – 2023

ЗМІСТ

ВСТУП

Актуальність теми дослідження

Штучний інтелект це сучасна технологія, яка постійно розвивається. Разом із цим, вона набуває все більшої популярності, розширює свої можливості та сфери застосування, полегшуючи людям життя. Вона ще далека до ідеалу, але невпинно до нього наближається.

Оскільки штучний інтелект покликаний певною мірою наслідувати людину, то і взаємодіяти з нею повинен як людина, тобто не лише текстовими даними, що вже було реалізовано, а й усним мовленням. Можливість вільно спілкуватись з комп’ютером може стати величезною економією часу та ресурсів.

Завдання та мета курсової роботи

Дослідити теоретичні аспекти розпізнавання природної мови штучним інтелектом, проблеми з якими зіштовхується, дослідити та порівняти види нейронних мереж, які застосовуються у цьому напрямку, а також розробити програму для наочної демонстрації його можливостей.

РОЗДІЛ 1. ПОНЯТТЯ ШТУЧНОГО ІНТЕЛЕКТУ ТА ОБРОБКИ ПРИРОДНОЇ МОВИ

1.1 Визначення штучного інтелекту

Штучний інтелект — це здатність інженерної системи до обробки, застосування та вдосконалення набутих знань та умінь. Хоч і точного визначення немає, оскільки сама природа людського інтелекту є нерозв'язаною темою в філософії, було запропоновано чимало гіпотез та підходів до розуміння задач штучного інтелекту.

Тест Тюринга, запропонований Аланом Тюрингом у 1950 році, був розроблений, щоб забезпечити те саме задовільне оперативне визначення інтелекту. Суть тесту полягала в наступному: комп'ютер проходить перевірку, якщо опитувач (людина), поставивши деякі письмові запитання, не зможе визначити від кого надходять письмові відповіді – від людини чи від комп'ютера.

В ході тесту було визначено, які можливості повинен мати комп'ютер для того, щоб пройти цей тест:

- обробка природної мови – для успішної комунікації з опитувачем;
- представлення знань – для збереження того, що він (комп'ютер) знає або чує;
- автоматизоване міркування – для використання збереженої інформації, щоб відповісти на запитання та дійти до нових висновків;
- машинне навчання – для адаптації до нових обставин, виявлення й застосування закономірностей;
- комп'ютерний зір – для сприйняття об'єктів;
- робототехніка – для маніпулювання об'єктами та пересування.

Останні два пункти необхідні для проходження повного тесту Тюринга, в якому комп'ютер повинен довести своє вміння усвідомлювати (розуміти) об'єкти через так звані органи чуття та взаємодіяти з ними.

1.2 Обробка природної мови

Обробка природної мови — це область розробки методів і алгоритмів, які отримують як вхідні або генерують як вихідні дані, відображені (записані) природною мовою. Як уже було зазначено вище, це один із основних розділів штучного інтелекту. Він спирається на багато інших інтелектуальних дисциплін: від формальної лінгвістики до статистичної фізики.

Головними завданнями цього напряму є:

- видобування даних (вивчення, зв'язки та закономірності);
- синтез мовлення (перетворення текстових даних у звуки, озвучення);
- розпізнавання мови (отримання текстових даних з картинок, відсканованих документів або з мовлення, продукованого людським голосом);
- генерування природної мови (перетворення комп'ютерних даних у людську мову);
- машинний переклад (переклад з однієї людської мови на іншу, без втрати змісту);
- питання-відповідь (відповіді на питання, задані природною мовою).

Розуміти та продукувати мову за допомогою комп'ютера дуже складно. І на це є багато причин, зокрема неоднозначність та постійне розширення мови. Наприклад, маємо два речення:

- Я їв піцу з родиною.
- Я їв піцу з грибами.

Значення першого не зміниться, якщо передати його як «Я з родиною їв піцу», тобто тут йдеться про групу людей, що споживають їжу. А ось в другому — про додаткові інгредієнти страви.

Розширенню мови сприяє постійний технологічний прогрес, поява нових продуктів, хвороб та просто розвиток мовлення. Також до цього можна віднести і можливі зміни у правописі, що теж ускладнюють завдання комп'ютеру.

РОЗДІЛ 2. РОЗПІЗНАВАННЯ ПРИРОДНОЇ МОВИ

2.1 Що таке розпізнавання природної мови

Природна мова часто передається в усній формі і розпізнавання мовлення — це завдання перетворення звукового сигналу в текст. З однієї точки зору, це проблема обробки сигналу, яку можна розглядати як етап попередньої обробки перед застосуванням обробки природної мови. Проте контекст відіграє вирішальну роль у розпізнаванні мовлення людьми. Саме з контексту люди можуть точніше визначити, про що їй каже співрозмовник, особливо, якщо серед сказаного присутні слова, що своїм звучанням схожі на інші. З цієї причини розпізнавання мовлення часто інтегрується з аналізом тексту, зокрема зі статистичними моделями мови, які кількісно визначають ймовірність послідовності тексту.

2.2 Автоматичне розпізнавання мовлення (АРМ)

Саме по собі розуміння розмовної мови є складним завданням, і дивно, що люди справляються з ним так добре. Метою дослідження автоматичного розпізнавання мовлення є вирішення цієї проблеми за допомогою обчислень, створюючи системи, які відображають акустичний сигнал у рядок слів. Автоматичне розуміння мовлення розширює цю мету до розуміння (певною мірою) всього речення, а не лише окремих слів.

Загальна проблема автоматичної транскрипції мовлення будь-яким мовцем у будь-якому середовищі ще далека від вирішення. Але за останні роки технологія АРМ дозріла до такого моменту, що стала життєздатною в певних обмежених областях. Однією з основних областей застосування є взаємодія людини та комп’ютера. Хоча багато завдань краще вирішуються за допомогою візуальних або вказівних інтерфейсів, мовлення може бути кращим інтерфейсом, ніж клавіатура. Наприклад для завдань, де повне спілкування природною мовою є корисним або для яких клавіатури не підходять. Сюди входять додатки, де зайняті руки або очі і

немає можливості використовувати клавіатуру, наприклад, де користувач має об’єкти для маніпулювання або обладнання для керування.

Іншою важливою сферою застосування є телефонна, де розпізнавання мовлення вже використовується, наприклад, у системах розмовного діалогу для введення цифр, розпізнавання «так» для прийому дзвінків, отримання інформації про літак або поїзд та маршрутизації дзвінків («З’єднайте мене із …»). У деяких програмах мультимодальний інтерфейс, що поєднує мовлення та інший вид діяльності, може бути ефективнішим, ніж графічний інтерфейс користувача без мови.

На останок, АРМ застосовується під час диктування, тобто транскрипції розширеного монологу одним певним мовцем. Диктант поширений у таких галузях, як право, а також важливий як частина комунікації доповнення (взаємодія між комп’ютерами та людьми з деякими вадами, що призвели до нездатності друкувати або говорити).

2.3 Складність та проблематика АРМ

Розпізнавання мовлення є досить складним, оскільки звуки, які видає мовець, неоднозначні та шумні. Навіть цього достатньо, щоб спричинити кілька проблем, які ускладнюють мовлення.

По-перше, сегментація — написані слова мають пробіли між собою, але у швидкому мовленні немає пауз (наприклад, англійською «wreck a nice» може звучати як одне слово «recognize»).

По-друге, коартикуляція — злиття звуків при швидкому мовленні (наприклад, звук «s» в кінці слова «nice» зливається зі звуком «b» на початку слова «beach», утворюючи щось, близьке до «sp»).

Vowels		Consonants B–N		Consonants P–Z	
Phone	Example	Phone	Example	Phone	Example
[iy]	beat	[b]	bet	[p]	pet
[ih]	bit	[ch]	Chet	[r]	rat
[eh]	bet	[d]	debt	[s]	set
[æ]	bat	[f]	fat	[sh]	shoe
[ah]	but	[g]	get	[t]	ten
[ao]	bought	[hh]	hat	[th]	thick
[ow]	boat	[hv]	high	[dh]	that
[uh]	book	[jh]	jet	[dx]	butter
[ey]	bait	[k]	kick	[v]	vet
[er]	Bert	[l]	let	[w]	wet
[ay]	buy	[el]	bottle	[wh]	which
[oy]	boy	[m]	met	[y]	yet
[axr]	diner	[em]	bottom	[z]	zoo
[aw]	down	[n]	net	[zh]	measure
[ax]	about	[en]	button		
[ix]	roses	[ng]	sing		
[aa]	cot	[eng]	washing	[-]	*silence*

Рис. 2.1 — Приклади транскрипцій букв у словах англійської мови

Ще однією проблемою є омофони (такі слова, як «to», «too» і «two», які звучать однаково, але відрізняються за значенням).

Оскільки набір усіх можливих речень будь-якої природної мови величезний, нам потрібен ефективний алгоритм, який не буде шукати всі можливі речення, а лише ті, які мають хороші шанси відповідати введеним. Це проблема декодування або пошуку. Оскільки простір пошуку настільки великий для розпізнавання мовлення, його ефективність є важливою частиною завдання, від цього залежить чи потрібне взагалі розпізнавання, адже якщо воно буде дуже повільним у ньому не буде сенсу.

Завдання, які може виконувати АРМ дуже багато і визначити їхню складність лише глянувши на них досить важко. Тому існують параметри, що допомагають у цьому.

Одним із вимірів варіації завдань на розпізнавання мовлення є обсяг словникового запасу. Зрозуміло, що розпізнавання мовлення є легшим, якщо кількість різних слів, які нам потрібно розпізнати, менша. Таким чином, завдання зі словником із двох слів (як-от «так» або «ні») або словником з десяти слів (як-от розпізнавання послідовностей цифр, що називається завданням «цифр») є відносно легкими. З іншого боку, завдання з великим словником (завдання зі словниковим запасом 64 000 слів і більше, як-от розшифрування телефонних розмов між людиною або транскрибування новин) набагато складніші.

Другим виміром варіації є те, наскільки вільною, природною чи розмовною є мова. Ізольоване розпізнавання слів, у якому кожне слово відокремлене певною паузою, є набагато простішим, ніж розпізнавання безперервного мовлення, в якому слова стикаються одне з одним. Самі завдання безперервного мовлення сильно відрізняються за складністю. Наприклад, мовлення людини до машини, як виявилось, набагато легше розпізнати, ніж розмову людини з людиною. Тобто розпізнати мовлення людей, які спілкуються з машинами, читають вголос або розмовляють зі системою мовного діалогу, відносно легко. А от розпізнати мовлення двох людей, які розмовляють одна з одною (наприклад, розшифрування ділової зустрічі чи телефонної розмови), набагато важче. Схоже, що коли люди розмовляють з машинами, вони дещо спрощують свою мову, розмовляючи повільніше і чіткіше.

Третій вимір варіації — канал і шум. Завдання диктанту (і багатьох лабораторних досліджень з розпізнавання мовлення) виконуються за допомогою високоякісних мікрофонів, встановлених на голові того, хто говорить. Вони призначені для усування спотворень, які виникають при використанні настільного мікрофона, коли голова мовця рухається. Будь-який шум також ускладнює розпізнавання. Таким чином, розпізнати мовлення мовця, що диктує в тихому офісі, набагато легше, ніж в галасливій машині на шосе з відкритим вікном.

Останнім виміром варіації є акцент або характеристики класу оратора. Мовлення легше розпізнати, якщо мовець говорить на стандартному діалекті або

взагалі на такому, який відповідає даним, на яких навчалася система. Таким чином, розпізнавання мови з іноземним акцентом або мовлення дітей (якщо система не була спеціально навчена саме цим видам мовлення) є важчим. Також до цього можемо віднести картавість та інші незначні відхилення у мовленні.

2.4 Прихована модель Маркова

Прихована модель Маркова (ПММ) — це тимчасова імовірнісна модель, в якій стан процесу описується однією дискретною випадковою величиною. Можливі значення змінної — це можливі стани світу.

Прихована модель Маркова дозволяє нам говорити як про спостережувані (наприклад, слова, які ми бачимо у вхідних даних), так і про приховані події (наприклад, теги частини мови), які ми вважаємо причинними факторами в нашій імовірнісній моделі.

ПММ у розпізнаванні мовлення використовується для визначення ймовірності певної послідовності літер або слів у реченні.

2.5 Модель шумового каналу

Завдання розпізнавання мовлення полягає в тому, щоб прийняти як вхідний сигнал акустичну форму хвилі і створити на виході текстовий рядок слів. Системи розпізнавання мовлення на основі прихованої моделі Маркова (ПММ) розглядають це завдання, використовуючи метафору шумового каналу. Інтуїція моделі шумового каналу полягає в тому, щоб розглядати форму акустичного сигналу як «зашумлену» версію рядка слів, тобто версію, яка була передана через шумовий канал зв’язку. Цей канал вносить «шум», який ускладнює розпізнавання «справжнього» рядка слів. Мета полягає в тому, щоб побудувати модель каналу, яка дозволить зрозуміти, як він змінив те «справжнє» речення, а отже, відновити його. Суть моделі шумового каналу полягає в тому, що якщо ми знаємо, як канал

спотворює джерело, ми могли б знайти правильне вихідне речення для форми сигналу, взявши всі можливі речення в мові, пропустивши кожне речення через нашу модель каналу з шумами та побачивши, чи він відповідає результату. Потім ми вибираємо найкраще відповідне вихідне речення як бажане вихідне речення.

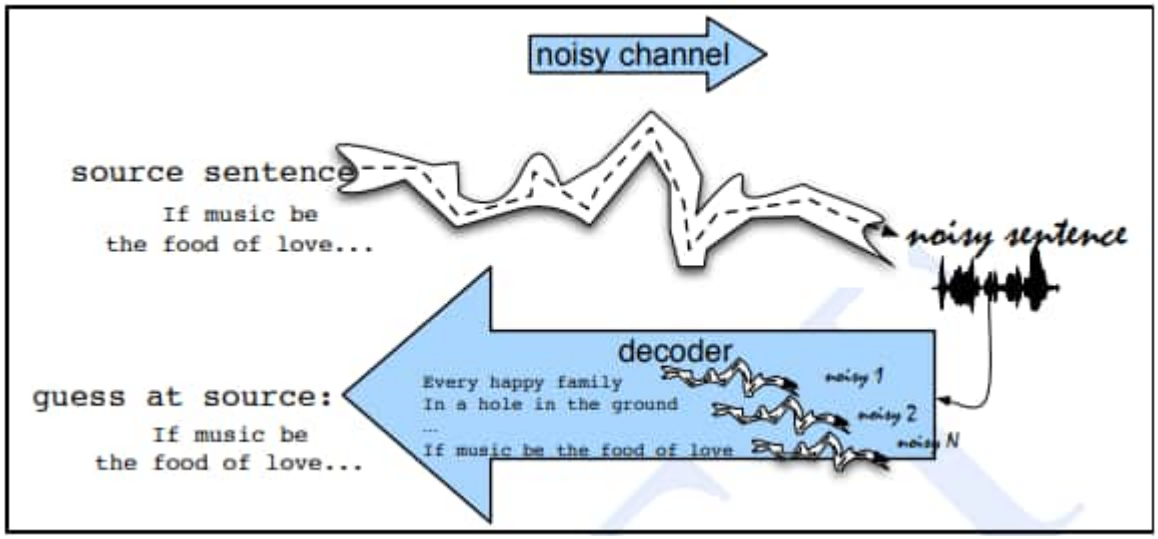

Рис. 2.2 — Схема моделі шумового каналу

Ми можемо розглядати розпізнавання мовлення як проблему в поясненні найімовірнішої послідовності. Тобто це проблема обчислення найбільш вірогідної послідовності змінних стану, з урахуванням послідовності спостережень. У цьому випадку змінними стану є слова, а спостереження — звуки. Точніше, спостереження — це вектор ознак, витягнутих із звукового сигналу. Як і зазвичай, найімовірнішу послідовність можна обчислити за допомогою правила Байєса:

$$\underset{word_{1:t}}{argmax}\, P(word_{1:t}|sound_{1:t}) = \underset{word_{1:t}}{argmax}\, P(sound_{1:t}|word_{1:t}) * P(word_{1:t}) \quad (2.1)$$

Тут $P(sound_{1:t}|word_{1:t})$ є акустичною моделлю. У ній описуються звуки слів. $P(word_{1:t})$ — мовна модель. Вона визначає попередню ймовірність кожного висловлювання. Наприклад, послідовність слів “ceiling fan” приблизно в 500 разів більш імовірна, ніж “sealing fan”.

Після визначення акустичної та мовної моделей ми можемо знайти найбільш вірогідну послідовність слів за допомогою алгоритму Вітербі. Більшість систем розпізнавання мовлення використовує мовну модель, яка робить припущення Маркова, яке полягає у тому, що поточний стан $Word_t$ залежить лише від фіксованої кількості n попередніх станів, і представляють $Word_t$ як одну випадкову величину, що приймає скінченний набір значень, що робить її Прихованою моделлю Маркова. Таким чином, розпізнавання мовлення стає простим застосуванням методології ПММ.

2.6 Акустична модель

Звукові хвилі — це періодичні зміни тиску, які поширюються в повітрі. Коли ці хвилі потрапляють на діафрагму мікрофона, рух вперед-назад створює електричний струм. Аналогово-цифровий перетворювач вимірює величину струму, через дискретні інтервали, які називаються частотою дискретизації. Звуки мовлення, які переважно знаходяться в діапазоні від 100 Гц (100 циклів в секунду) до 1000 Гц, зазвичай відбираються з частотою 8 кГц.

Розпізнавачі мовлення зазвичай зберігають від 8 до 12 біт. Це означає, що система низького класу, з дискретизацією на частоті 8 кГц з 8-бітним квантуванням, вимагала б майже пів мегабайта на хвилину мовлення. Оскільки ми хочемо знати лише те, які слова були сказані, а не як вони звучали, то немає потреби зберігати всю цю інформацію. Треба лише розрізняти різні звуки мовлення. Лінгвісти виявили близько 100 звуків мови, які можна скласти, щоб утворити всі слова у всіх відомих людських мовах. Грубо кажучи, це звуки, кожен з яких відповідає одному голосному чи приголосному, але є деякі труднощі: комбінації літер, наприклад «th» і «ng», утворюють окремі звуки, а деякі букви утворюють різні звуки в різних ситуаціях (наприклад, в англійській мові словосполучення “Pacific Ocean” містить в собі 3 літери «c» і кожна звучатиме по-різному).

Фонема — це найменша звукова одиниця, яка має чітке значення для носіїв певної мови. Наприклад, «t» у слові «stick» звучить досить схоже на «t» у «tick», що носії англійської мови вважають їх однією і тією ж фонемою. Але різниця суттєва в тайській мові, тому там дві фонеми. Щоб представляти розмовну англійську, нам потрібно уявлення, яке може розрізняти різні фонеми, але таке, яка не потребує розрізнення нефонематичних варіацій звуку: гучний чи тихий, швидкий чи повільний, чоловічий чи жіночий голос тощо.

По-перше, ми помічаємо, що хоча звукові частоти в мові можуть становити кілька кГц, зміни в змісті сигналу зустрічається набагато рідше, можливо, не більше 100 Гц. Тому мовні системи підсумовують властивості сигналу в часових зрізах, які називаються кадрами. Довжина кадру близько 10 мілісекунд є достатньо короткою, щоб гарантувати, що деякі короткотривалі явища будуть упущені. Перекриваючі кадри використовуються, щоб переконатися, що ми не пропустимо сигнал, оскільки він потрапляє на межу кадру. Кожен кадр підсумовується вектором ознак. Виділити ознаки з мовного сигналу — це по-суті те саме, що слухати оркестр і виділяти кожен інструмент на фоні інших.

Короткий огляд функцій типової системи: спочатку використовується перетворення Фур’є для визначення кількості акустичної енергії приблизно на десятку частот. Потім ми обчислюємо міру, яка називається кепстральним коефіцієнтом частоти або MFCC для кожної частоти. Ми також обчислюємо повну енергію в кадрі. Це дає тринадцять ознак; для кожного з них ми обчислюємо різницю між цим кадром і попереднім кадром, а також різницю між відмінностями, загалом — 39 ознак. Це неперервні значення; найпростіший спосіб вмістити їх у структуру ПММ — дискретизувати.

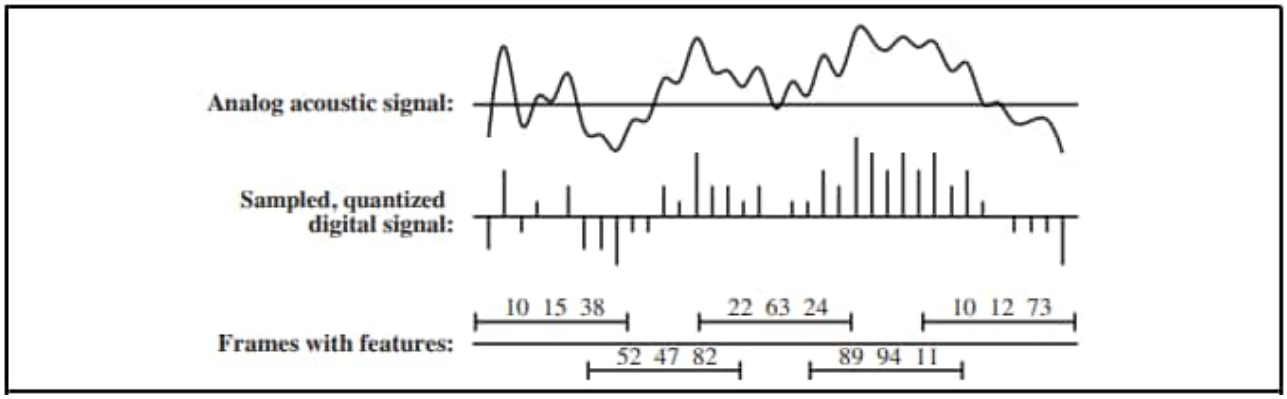

Рис. 2.3 — Перетворення акустичного сигналу в послідовність кадрів

Після того як від необробленого акустичного сигналу ми перейшли до серії спостережень, необхідно описати (неспостережувані) стани ПММ і визначити модель переходу $P(X_t \mid X_{t-1})$ і модель датчика $P(E_t \mid X_t)$. Модель переходу може бути розбита на два рівні: слово і фонема. Почнемо знизу: модель фонеми описує фонему у трьох станах: початок, середина та кінець. Наприклад, звук [t] має тихий початок, невеликий вибуховий сплеск звуку в середині та (зазвичай) шипіння в кінці. Модель фонеми також може описувати звучання цілого слова, а не лише окремих фонем.

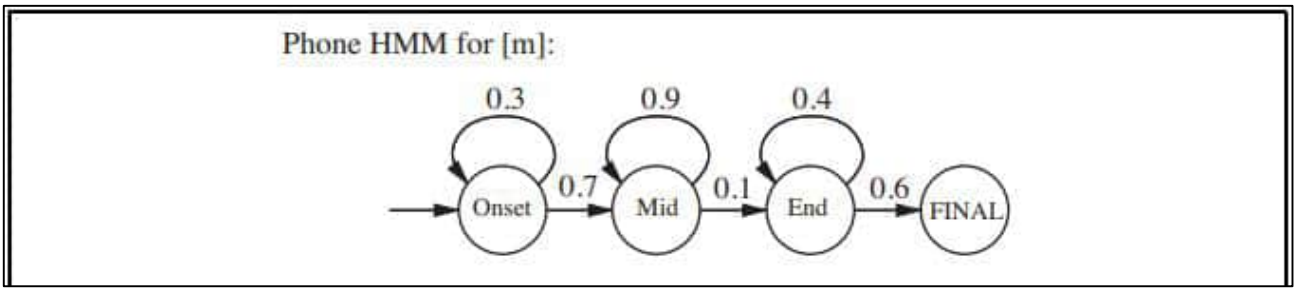

Рис. 2.4 — Приклад моделі фонеми [m] з використанням Прихованої моделі Маркова

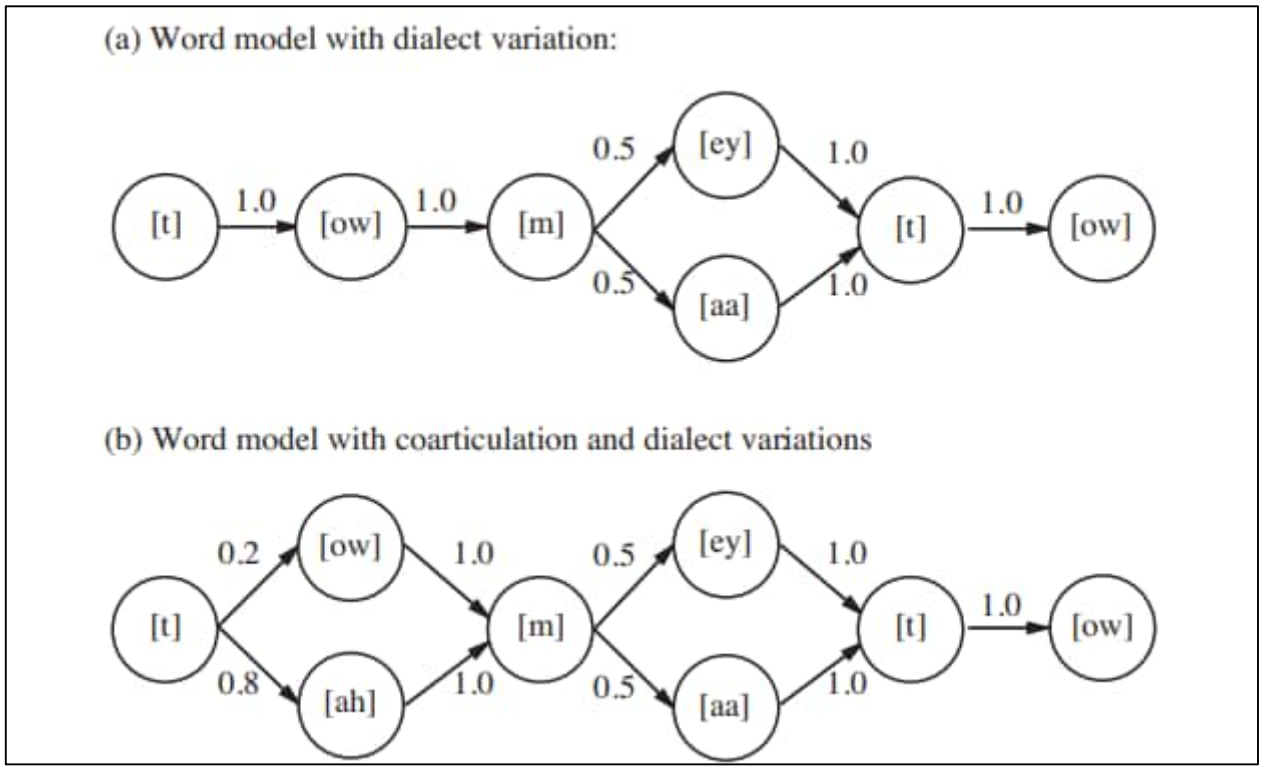

Рис. 2.5 — Приклад моделі вимови з використанням Прихованої моделі Маркова, де присутні варіації вимови слова внаслідок акценту

2.7 Мовна модель

Для розпізнавання мовлення загального призначення мовна модель може бути n-грамною моделлю тексту, засвоєного з корпусу письмових речень. Однак слід пам’ятати, що розмовна мова дещо відрізняється від письмової, тому краще використовувати корпуси транскриптів розмовної мови. Для розпізнавання мовлення для конкретного завдання корпуси теж повинні бути конкретними (наприклад, щоб створити систему бронювання авіакомпаній, отримайте стенограми попередніх дзвінків). Це також допомагає отримати словниковий запас для конкретних завдань (наприклад, список усіх аеропортів і міст, які обслуговуються, а також номери всіх рейсів). Частина дизайну голосового інтерфейсу користувача полягає в тому, щоб примушувати користувача говорити речі за допомогою обмеженого набору параметрів, щоб розпізнавач мовлення мав більш жорсткий розподіл ймовірностей. Наприклад, на запитання «У яке місто Ви

хочете поїхати?» очікується відповідь із дуже обмеженою мовною моделлю. З іншого боку, запитання «Чим я можу вам допомогти?» аж ніяк.

2.8 Побудова розпізнавача мовлення

Якість системи розпізнавання мовлення залежить від якості всіх її компонентів — мовної моделі, моделей вимови слів, моделей фонем та алгоритмів обробки сигналів, які використовуються для вилучення спектральних характеристик з акустичного сигналу.

Структура моделей вимови, наприклад, моделі слова «томат» на рисунку 2.5, зазвичай розробляється вручну. Великі словники вимови тепер доступні для англійської та інших мов, хоча їх точність дуже різниться. Структура моделей фонем із трьома станами однакова для всіх фонем, як показано на рисунку 2.4. Відрізняються лише ймовірності. Їх можна отримати з корпусу мовлення. Найпоширенішим типом корпусу є той, який включає мовленнєвий сигнал для кожного речення в парі з розшифруванням слів. Побудувати модель з цього корпусу складніше, ніж побудувати n-грамову модель тексту, тому що ми повинні побудувати приховану модель Маркова — послідовність фонем для кожного слова і їх стан для кожного періоду часу є прихованими змінними. У перші дні розпізнавання мови приховані змінні забезпечувалися трудомістким ручним маркуванням спектрограм. Однак сучасні системи використовують максимізацію очікування для автоматичного надання відсутніх даних. Ідея проста: враховуючи ПММ і послідовність спостереження, ми можемо використовувати алгоритми згладжування для обчислення ймовірності кожного стану на кожному кроці часу і, шляхом простого розширення, ймовірність кожної пари стан-стан на послідовних кроках часу. Ці ймовірності можна розглядати як невизначені мітки. З невизначених міток ми можемо оцінити ймовірність нових переходів і датчиків, і процедура максимізації очікування повторюється. Цей метод гарантовано збільшує відповідність між моделлю та даними на кожній ітерації, і, як правило, він

сходить до набагато кращого набору значень параметрів, ніж ті, які надаються початковими оцінками, позначеними вручну.

Системи з найвищою точністю працюють, навчаючи різні моделі для кожного мовця, таким чином вловлюючи відмінності в діалекті, а також чоловічі/жіночі та інші варіації. Для цього навчання може знадобитися кілька годин взаємодії з оратором, тому системи, які мають найбільше поширення, не створюють специфічних для оратора моделей.

Точність системи залежить від ряду факторів. По-перше, має значення якість сигналу: високоякісний мікрофон, спрямований на нерухомий рот мовця у кімнаті з м’якими стінами, буде працювати набагато краще, ніж дешевий мікрофон, який передає сигнал по телефонних лініях від автомобіля, який перебуває в заторі з увімкненим радіо. Розмір словникового запасу теж має значення: при розпізнаванні зі словниковим запасом в 11 слів частота помилок буде нижчою за 0,5%, тоді як у новинах де словниковий запас містить близько 20 000 слів вона зросте до 10% і до 20% на корпус із 64 000 слів. Завдання також має значення: коли система намагається виконати конкретне завдання — забронювати рейс або дати маршрут до ресторану — завдання часто можна виконати ідеально, навіть якщо рівень помилок слів становить 10% і більше.

РОЗДІЛ 3. ВИДИ НЕЙРОННИХ МЕРЕЖ, ЩО ВИКОРИСТОВУЮТЬСЯ У РОЗПІЗНАВАННІ ПРИРОДНОЇ МОВИ

3.1 Згорткова нейронна мережа (ЗНМ)

3.1.1 Нейронні мережі прямого поширення

Усі нейронні мережі поділяються на види (класи) залежно від принципів їхньої роботи. Одним із таких класів є нейронні мережі прямого поширення (рисунок 3.1). У них сигнали поширюються в одному напрямку, починаючи від шару вхідних нейронів *x*, через приховані шари *h* до вихідного шару *y* і на вихідних нейронах отримується результат опрацювання сигналу.

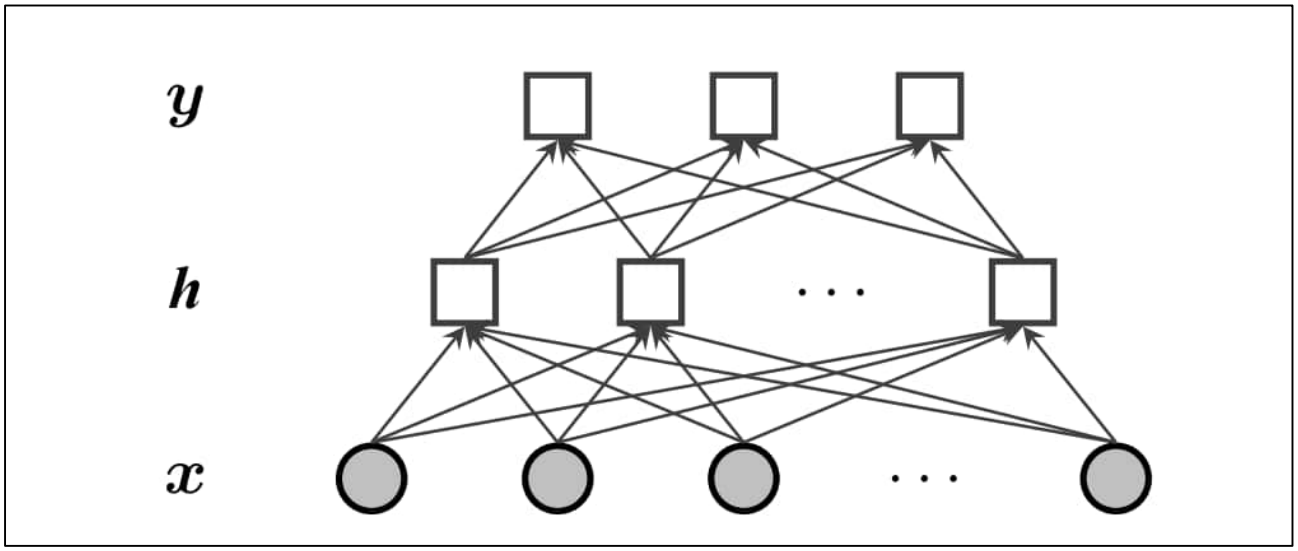

Рис. 3.1 — Схема нейронної мережі прямого поширення

3.1.2 Поняття та архітектура ЗНМ

Згорткові мережі беруть свій початок від невролога Девіда Г'юбела та нейробіолога Торстена Візела, які під час вивчення нейронів, які використовуються для локальної чутливості та вибору орієнтації в корі головного мозку котів, виявили, що єдина мережева архітектура може зменшити складність нейронної мережі.

ЗНМ часто використовуються в обробці зображень. А для їх застосування в розпізнаванні природної мови акустичні сигнали перетворюють на спектрограми (рисунок 3.2), з якими мережа вже може працювати.

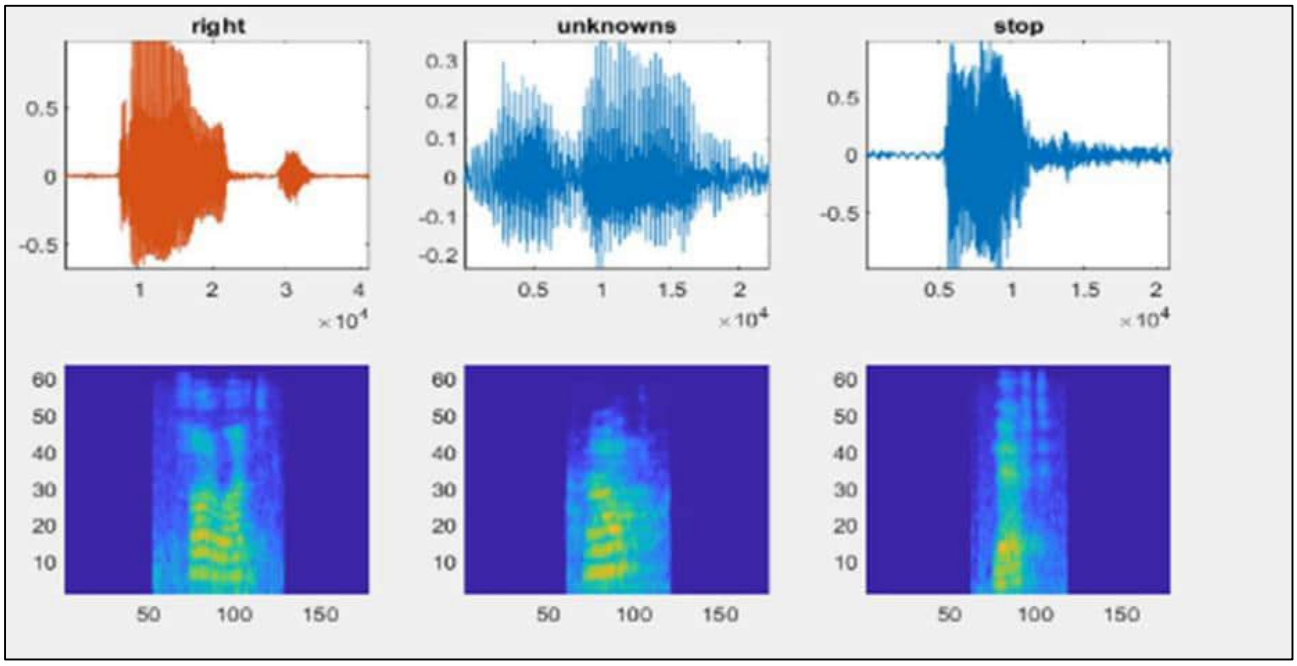

Рис. 3.2 — Приклади акустичних сигналів та спектрограм, створених на їхній основі

ЗНМ має фільтр, який переміщує зображення до створеної карти функцій на шарах згортки. Через це вікно або фільтр ваги мережі можуть ідентифікувати різні характеристики вхідного зображення. Функція активації вирішує, чи є певна функція в певному місці на зображенні. Зазвичай над зображенням застосовують багато фільтрів, щоб знайти необхідні функції.

Архітектура цієї мережі має три основні ідеї:

- локальність (здатність зменшувати вплив небілого шуму);
- розподіл ваги (покращує міцність моделі, зменшує перенавченість, зменшує кількість ваг);
- об’єднання (зменшує розмір).

Згорткову мережу часто називають локальною, оскільки окремі блоки, обчислені в певному місці вікна, залежать від локальної області, на яку зараз

«дивиться» вікно. Архітектура координується трьома основними шарами, розташованими в структурі прямого поширення: згортковий шар (для виділення ознак), шар підвибірки та шар об’єднання (для зменшення розмірів вхідних і вихідних даних). Також присутні лінійний фільтр і нелінійна функція активації — одні з найважливіших елементів.

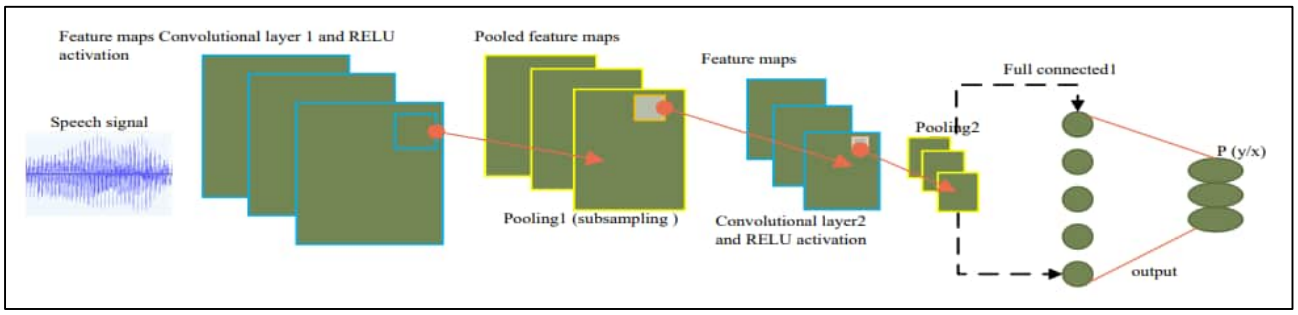

Рис. 3.3 — Архітектура шарів ЗНМ

У згортковому шарі (Convolutional layer) кожна площина з’єднана з однією або декількома картами функцій (Feature maps) попереднього шару. Функція активації застосовується до результату, який отримує вихід площини. Площинним виходом є двовимірна матриця — та ж карта функцій, назва якої пішла від того, що кожен результат згортки вказує на присутність візуальної функції в конкретному місці пікселя. Шар згортки створює одну або більше карт функцій. Потім кожна з них з’єднується точно з однією площиною в наступному шарі об’єднання (Pooling).

Спільне використання ваг і розташування мають важливе значення для властивостей об’єднання. Значення функцій, обчислених в різних місцях, групуються разом і представляються одним значенням, щоб мінімізувати відмінності у витягнутих функціях уздовж частотного виміру, коли вхідні моделі зсуваються. Це дозволяє працювати з невеликими частотними зсувами, поширеними в мові, що є наслідком різної довжини вокалу в людей.

Розглянемо приклад. Припустимо, що після перетворення акустичного сигналу ми отримали спектрограму, яку представлено матрицею розміру 6*6, крім того, у нас є фільтр (матриця 3*3). Виконавши згортання, ми отримаємо матрицю 4*4 (6 – 3 + 1 = 4), яку можемо побачити на рисунку 3.4.

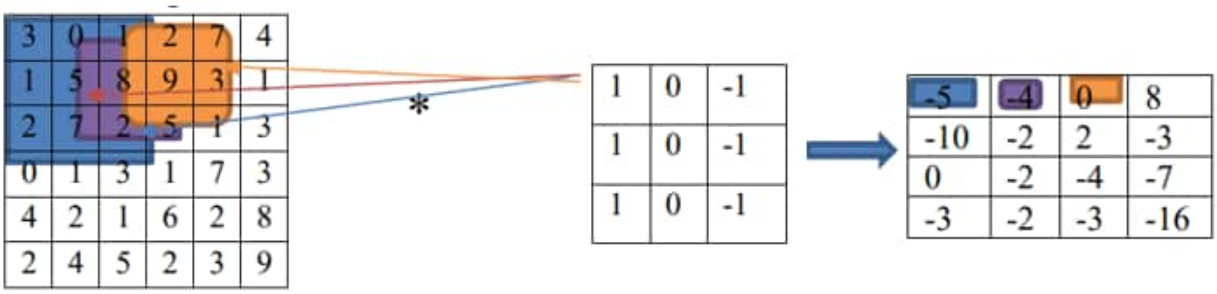

Рис. 3.4 — Процес згортання

3.2 Рекурентна нейронна мережа (РНМ)

3.2.1 Поняття РНМ

Рекурентною нейронною мережею (РНМ) називають будь-яку нейронну мережу, яка містить цикл у своїх мережевих з'єднаннях. Тобто будь-яка мережа, де кожне значення прямо чи опосередковано залежить від попередніх вихідних даних, які для поточного значення будуть сприйматися як вхідні. Хоча такі мережі є потужними, їх важко навчити. Однак у загальному класі рекурентних мереж існують обмежені архітектури, які виявилися надзвичайно ефективними при застосуванні до усної мови. До таких належать мережі з довгою короткочасною пам'яттю.

3.2.2 Проста РНМ

Проста рекурентна нейронна мережа — один із видів РНМ, яку ще називають мережею Елмана.

Розглянемо рисунок 3.5, на якому зображена спрощена схема простої РНМ. Як і у звичайних мережах прямого зв'язку, вхідний вектор, що представляє поточний вхідний елемент, x_t, множиться на вагову матрицю, а потім передається через функцію активації для обчислення значення активації для шару прихованих значень h_t. Цей прихований шар, у свою чергу, використовується для обчислення відповідного виходу y_t. У цьому випадку послідовності обробляються шляхом подання в мережу одного елемента за раз. Ключова відмінність від прямої мережі

полягає в повторюваному зв’язку, показаному на рисунку пунктирною лінією. Це доповнює вхідні дані, з якими будуть проведені обчислення у прихованому шарі, значенням активації прихованого шару з попереднього моменту часу.

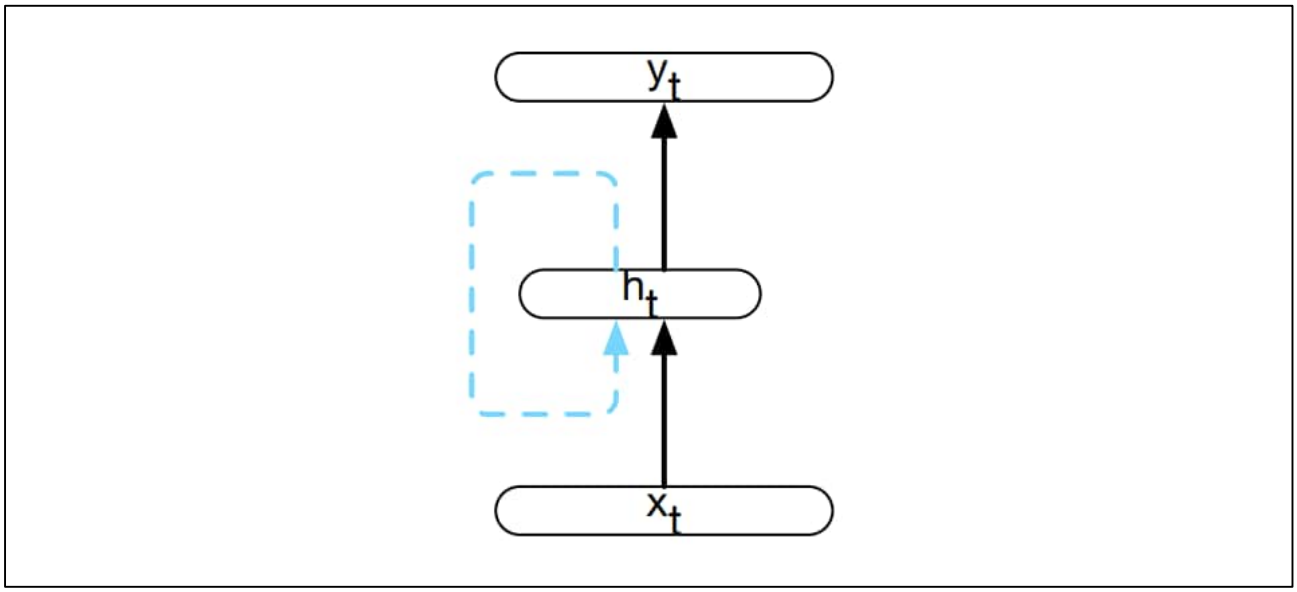

Рис. 3.5 — Схема простої РНМ

Прихований шар із попереднього часового кроку забезпечує форму пам’яті або контексту, який закодовує попередню обробку та інформує про рішення, які мають бути прийняті в наступні моменти часу. Важливим є те, що ця архітектура не накладає обмеження фіксованої довжини на попередній контекст; контекст, створений у попередньому прихованому шарі, містить у собі інформацію, що тягнеться від початку всієї послідовності.

Додавання такого тимчасового виміру може зробити рекурентні нейронні мережі більш екзотичними за решту архітектур. Але насправді вони не такі вже й різні. Враховуючи вхідний вектор і значення для прихованого шару з попереднього часового кроку, ми все ще виконуємо стандартне пряме обчислення. Щоб побачити це, розглянемо рисунок 3.6, який пояснює природу повторення та те, як воно впливає на обчислення на прихованому рівні. Найсуттєвіша зміна полягає в новому наборі вагових коефіцієнтів U, які з’єднують прихований шар із попереднього часового кроку з поточним прихованим шаром. Ці ваги визначають, як мережа повинна використовувати минулий контекст для обчислення виходу для поточного

введення. Як і з іншими вагами в мережі, ці з'єднання навчаються за допомогою зворотного поширення.

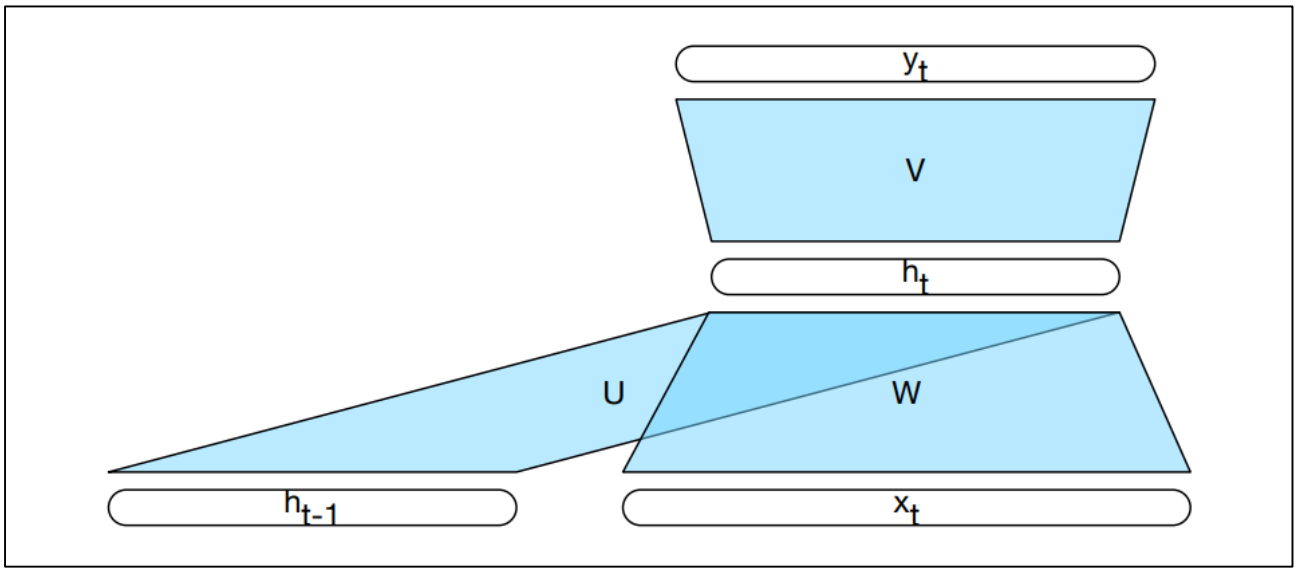

Рис. 3.6 — Проста РНМ, зображена у вигляді мережі прямого зв'язку

Прямий висновок (відображення послідовності входів у послідовність виходів) у РНМ майже ідентичний тому, що ми вже бачили з мережами прямого зв'язку. Щоб обчислити вихід y_t для входу x_t, нам потрібно значення активації для прихованого шару h_t. Щоб обчислити його, ми множимо вхідні дані x_t на вагову матрицю W, а прихований шар із попереднього часового кроку h_{t-1} на вагову матрицю U. Ми додаємо ці значення і пропускаємо їх через відповідну функцію активації g, щоб отримати значення активації для поточного прихованого шару h_t. Отримавши значення, ми переходимо до звичайних обчислень для генерації вихідного вектора.

Формулою це можна зобразити так:

$$h_t = g(U * h_{t-1} + W * x_t) \tag{3.1}$$

$$y_t = f(V * h_t) \tag{3.2}$$

У звичайному випадку м'якої класифікації обчислення y_t складається з обчислення м'якого максимуму, яке забезпечує нормалізований розподіл ймовірностей за можливими класами виходу:

$$y_t = softmax(V * h_t) \tag{3.3}$$

Той факт, що для обчислення в момент часу t потрібне значення прихованого шару з моменту часу t-1, вимагає алгоритму інкрементального висновку, який продовжується від початку послідовності до кінця. Послідовну природу простих рекурентних мереж також можна побачити, розгорнувши мережу на часовій прямій, як показано на рисунку 3.7. На ньому зображено, як різні шари копіюються для кожної ітерації, щоб проілюструвати, що вони матимуть різні значення залежно від часу. Однак вагові матриці залишаються незмінними.

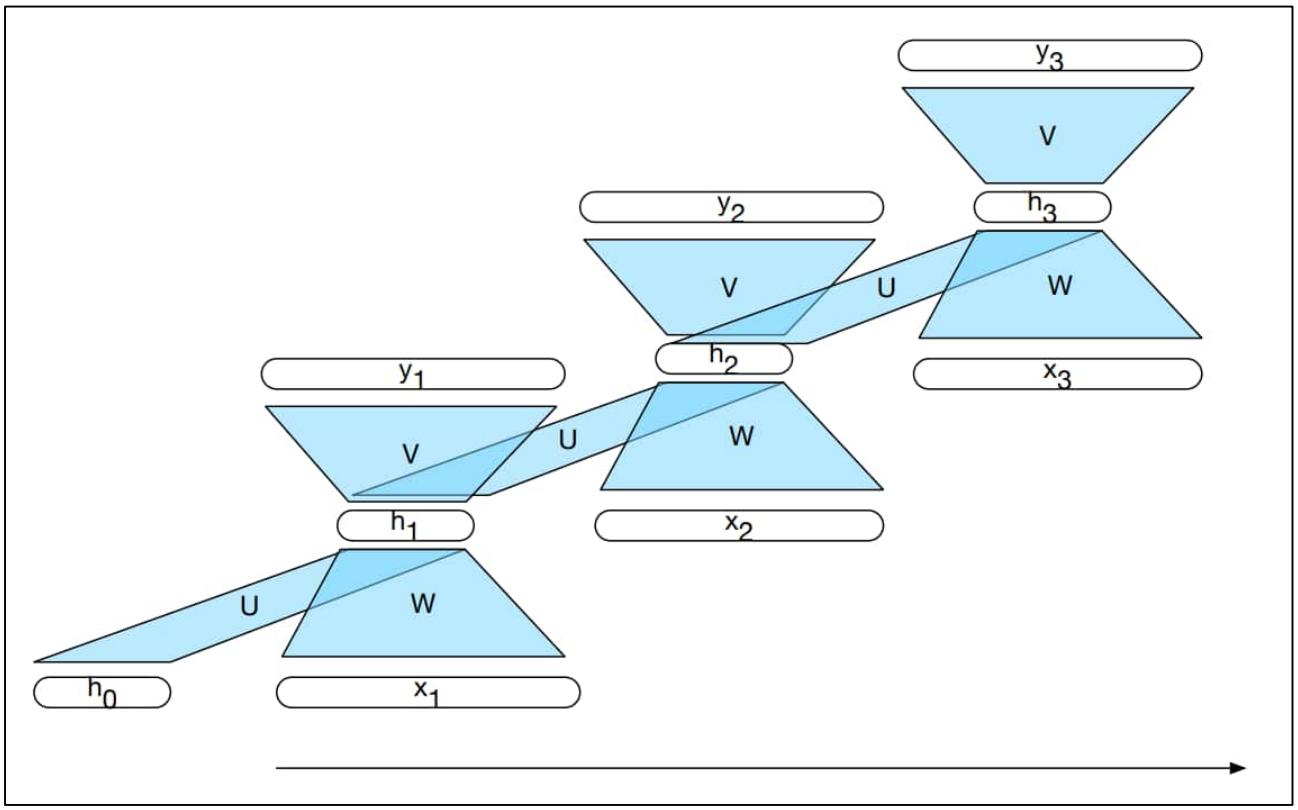

Рис. 3.7 — Схема простої РНМ, розгорнутої на часовій прямій

3.2.3 Мережі довгої короткочасної пам’яті (ДКЧП)

Проста РНМ є досить ефективною, але в неї є чимало недоліків. Головний із них — зменшення впливу по віддаленню. Тобто, на поточну ітерацію найбільше впливу має попередня, а решта, що були до неї, все менше і менше беруть участь у прогнозуванні. Це є серйозною проблемою, особливо у випадках, коли важлива для конкретної ітерації інформація знаходиться більш ніж за кілька кроків.

Вирішити таку проблему можуть мережі довгої короткочасної пам’яті. Вони були запропоновані 1997 року Зеппом Хохрайтером та Юрґеном Шмідгубером. ДКЧП розділяють проблему управління контекстом на дві підпроблеми: видалення з контексту інформації, яка більше не є потрібною, і додавання інформації, яка, ймовірно, знадобиться для подальшого прийняття рішень. Ключ до вирішення обох проблем полягає в тому, щоб навчитися керувати цим контекстом, а не просто вписувати стратегію в архітектуру. ДКЧП досягають цього, спочатку додаючи явний контекстний шар до архітектури (на додаток до звичайного повторюваного прихованого шару), а також за допомогою використання спеціалізованих нейронних блоків, які використовують «шлюзи» (рисунок 3.8) для керування потоком інформації, що надходить і виходить з блоків, з яких складаються мережеві шари. Ці шлюзи реалізуються за допомогою використання додаткових ваг, які діють послідовно на вхідному, попередньому прихованому шарі та попередньому контекстному шарі.

«Шлюзи» в ДКЧП мають загальний дизайн: кожен складається з прямого шару, за яким слідує сигмоїдна функція активації, а потім поточкове множення із шаром, який пройшов через «шлюз». Вибір сигмоїда як функції активації виникає через його тенденцію підштовхувати свої виходи до 0 або 1. Поєднання цього з поточковим множенням має ефект, подібний до ефекту двійкової маски. Значення в шарі, що проходить через «шлюз», які близькі за значеннями до 1, пропускаються майже без змін; значення, що ближчі до 0, по суті стираються.

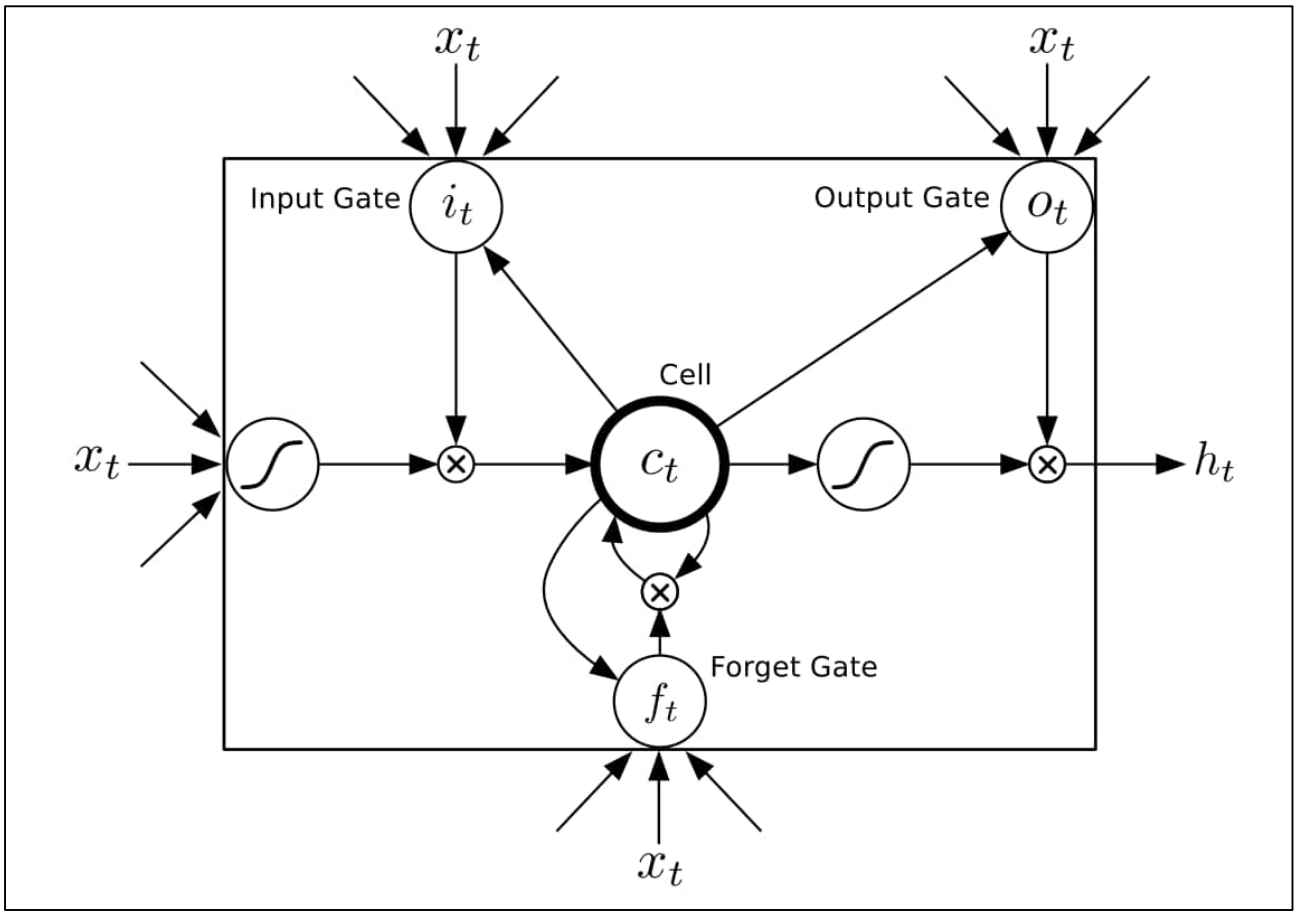

Рис. 3.8 — Схема «шлюзів» ДКЧП

Розглянемо «шлюз забуття» (Forget Gate). Його мета — видалити з контексту інформацію, яка більше не потрібна. «Шлюз забуття» обчислює зважену суму прихованого шару попередньої ітерації та поточного входу та пропускає її через сигмоїд. Ця маска потім множиться на вектор контексту, щоб видалити непотрібну інформацію.

Наступним завданням є обчислення фактичної інформації, яку нам потрібно витягти з попереднього прихованого шару та поточних вхідних даних шляхом їх додавання. Тоді створюється маска для того, щоб обрати інформацію для додавання до поточного контексту.

Останнім «шлюзом», який ми будемо використовувати, є «шлюз виходу», який використовується, щоб вирішити, яка інформація потрібна для поточного прихованого стану. Решта інформації зберігається для майбутніх рішень.

3.3 Порівняння ЗНМ та РНМ у розпізнаванні природної мови

Згорткова нейронна мережа та рекурентна нейронна мережа представляють протилежні класи нейромереж. У них різні принципи та підходи до виконання завдань. Однак вони обидві використовуються для розпізнавання природної мови.

Порівнявши ЗНМ та РНМ, очевидним стає факт переваги саме згорткової мережі. В першу чергу це зумовлено тим, що загалом мережі прямого поширення краще виконують завдання, пов’язані з розпізнаванням мовлення. Рекурентні мережі хоч і потужні, але поступаються ЗНМ у точності (85-90% проти 95% і вище) та є важкими в плані навчання.

РОЗДІЛ 4. ПРАКТИЧНИЙ ПРИКЛАД ВИКОРИСТАННЯ РОЗПІЗНАВАННЯ ПРИРОДНОЇ МОВИ ЯК ІНТЕРФЕЙСУ

4.1 Коротко про програму

Написана мною програма демонструє можливості усного мовлення в якості інтерфейсу за допомогою бібліотек, що використовують розпізнавання природної мови. У програмі наведено такі приклади використання: голосове меню, тестування та оператор бронювання авіарейсів.

4.2 Голосове меню

Вибір прикладу для демонстрації обирається голосом, при цьому вказувати опцію повністю не потрібно, а лише одне слово (тестування, оператор, бронювання, авіарейсів)

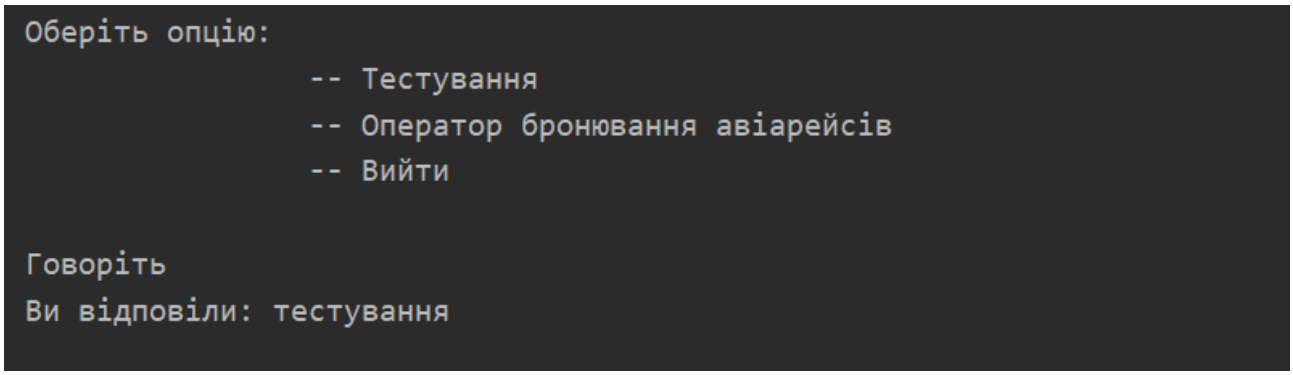

Рис. 4.1 — Приклад роботи голосового меню

4.3 Тестування

При виборі опції «Тестування» користувач отримує кілька питань та відповідає на них, називаючи відповіді, та отримує оцінку. Такий вид опитування може зекономити час на запис відповіді у текстовому чи іншому форматі.

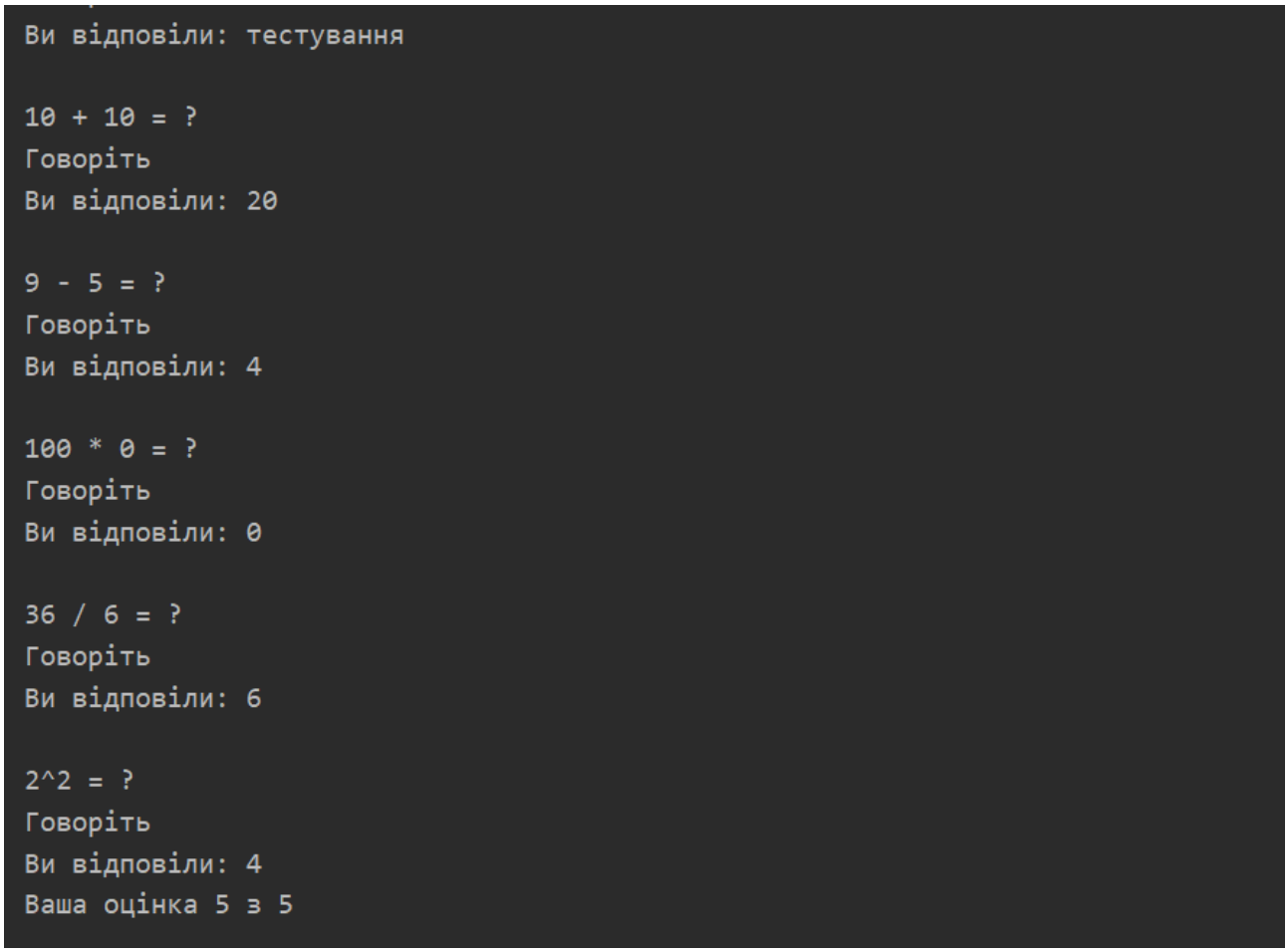

Рис. 4.2 — Приклад роботи тестування

4.4 Оператор бронювання авіарейсів

Ще одним прикладом використання усного мовлення як інтерфейсу може бути робота оператора, чиїм завданням є відповісти клієнту, чи існує авіарейс з міста А в місто Б.

```
Ви відповіли: оператор

Назвіть місто, з якого плануєте летіти
Говоріть
Ви відповіли: львів

Назвіть місто, до якого плануєте летіти
Говоріть
Ви відповіли: київ

Авіаперельотів між містами Львів та Київ зараз немає. Бажаєте продовжити?
Говоріть
Ви відповіли: ні
```

Рис. 4.3 — Приклад роботи оператора бронювання авіарейсів

ВИСНОВКИ

У своїй курсовій роботі я коротко оглянув саме поняття штучного інтелекту (його завдання та необхідні можливості) та його розділу про обробку природної мови (суть та приклади труднощів) для того, щоб більшою мірою розкрити головну тему — розпізнавання природної мови. Описав головні перешкоди, що стоять на заваді створення повноцінного штучного інтелекту, який зміг би повністю розуміти людину. Навів приклади про сучасні можливості, написав про засоби (моделі), що дозволяють досягти найкращого результату в даній сфері. Дослідив згорткову та рекурентну нейронні мережі, порівняв їх, віддавши перевагу першій. Також я написав програму для наочної демонстрації можливостей цього напряму.

ДОДАТОК

Під час виконання практичної частини було використано:

- Python – високорівнева, динамічно типізована, інтерпретована об'єктно-орієнтована мова програмування.
- SpeechRecognition – бібліотека для виконання розпізнавання мовлення.
- gTTS – бібліотека від Google для виконання розпізнавання мовлення.
- PyAudio – бібліотека для роботи зі звуком/аудіо.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Artificial Intelligence: A Modern Approach, 3rd ed / Stuart J. Russell and Peter Norvig. / 2010, 2003, 1995 by Pearson Education, Inc., Upper Saddle River, New Jersey 07458. – 1151 c. – ISBN-13: 978-0-13-604259-4.
2. Neural Network Methods for Natural Language Processing / Yoav Goldberg. / 2017 by Morgan & Claypool. – 309 c. - ISBN: 9781627052955.
3. Natural Language Processing / Jacob Eisenstein / November 13, 2018.
4. Speech and Language Processing / Daniel Jurafsky, James H. Martin. / 2008 by Prentice-Hall, Inc. Pearson Higher Education, Upper Saddle River, New Jersey 07458. – 1038 c. – ISBN 0-13-095069-6.
5. Speech Recognition using Convolution Deep Neural Networks / Ayad Alsobhani, Hanaa M A ALabboodi, Haider Mahdi / Journal of Physics: Conference Series 1973 (2021) 012166 doi:10.1088/1742-6596/1973/1/012166