НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ "КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО"

Факультет електроніки

(повна назва інституту/факультету)

Акустичних та мультимедійних електронних систем

(повна назва кафедри)

«До захисту допущено»

Завідувач кафедри

_____ С. А. Найда

«_10__» _06_____ 2022 р.

Дипломна робота

на здобуття ступеня бакалавра

зі спеціальності (спеціалізації) 171 Електроніка (Електронні системи мультимедіа та засоби інтернету речей)

на тему: Моделі мовленнєвих сигналів та методи їх цифрової обробки.

Виконав: студент IV курсу, групи ДВ-81

Богоденко Дмитро Олександрович

Керівник д.т.н., професор Розорінов Г.М. _____

Рецензент д.т.н., доц. кафедри МЕ Татарчук Д.Д. _____

Засвідчую, що в цій дипломній роботі немає запозичень із праць інших авторів без відповідних посилань.

Національний технічний університет України
«Київський політехнічний інститут
імені Ігоря Сікорського»

Факультет Електроніки

Кафедра акустичних та мультимедійних електронних систем

Рівень вищої освіти — перший (бакалаврський)

Спеціальність 171 Електроніка (Електронні системи мультимедіа та засоби інтернету речей)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Сергій Найда
«_02» _____05_____ 2022 р.

ЗАВДАННЯ
на дипломну роботу студенту
Богоденку Дмитру Олександровичу

1. Тема роботи: «Моделі мовленнєвих сигналів та методи їх цифрової обробки.»,керівник роботи: Розорінов Георгій Миколайович, професор
затверджені наказом по університету від «06» червня 2022 р. № 911-С.
2. Термін подання студентом роботи: «14» червня 2022 р.
3. Вихідні дані до роботи: системи розпізнавання мови, моделі синтезу мови, перетворення мовного сиганлу в цифрову форму
4. Зміст роботи: структурувати понятєвий апарат галузі обробки мовленнєвих сигналів, узагальнити теоретичне і практичне значення технологій розпізнавання мови
5. Перелік ілюстративного матеріалу: набір слайдів презентації з узагальнюючими матеріалами та описом створеного ефекту.

6. Консультанти розділів роботи:

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання: «08» лютого 2022 р.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Оформлення першого розділу роботи	16.05.2022	виконано
2	Оформлення другого розділу роботи	23.05.2022	виконано
3	Оформлення третього розділу роботи	9.06.2022	виконано
4	Підготовка матеріалів до друку та оформлення пояснювальної записки	01.06.2022	виконано
5	Підготовка та оформлення презентації для доповіді	10.06.2022	виконано

Студент  Богоденко Д. О.

Керівник роботи _____ Розорінов Г.М.

УДК 621.397.6

РЕФЕРАТ

Богоденко Д.О.. Моделі мовленнєвих сигналів та методи їх цифрової обробки: дипломна робота бакалавра : 171 Електроніка. – Київ, 2022. – 87 с.

Дипломну роботу виконано на 87 аркушах, вона містить 1 додаток та перелік посилань на використані джерела з 44 найменувань. У роботі наведено 22 рисунка Ключові слова: МОВЛЕННЄВІ СИГНАЛИ, РОЗПІЗНАВАННЯ МОВИ, СИСТЕМИ РОЗПІЗНАВАННЯ.

Мета роботи: аналіз технологій розпізнавання мовлення та дослідження їх точності та ефективності для покращення показників.

Об'єкт дослідження: Системи розпізнавання мови та їх функціональність у використанні .

Предмет дослідження: точність розпізнавання мовлення та функціональність систем.

Актуальність теми бакалаврської роботи полягає у вивченні систем розпізнавання мовлення та їх удосконалення для поширення в суспільстві для широкого кола завдань

Методи дослідження. В дослідницькій роботі використовувалися методи *теоретичного синтезу* для роботи з дослідницькою літературою та існуючими проблемами в галузі розпізнавання мови; метод *наукового узагальнення* для підготовки висновків в кінці кожного розділу і в кінці роботи для виокремлення чинників бакалаврського дослідження.

Наукова новизна полягає у винайденні новітніх рішень для поліпшення точності,ефективності та надійності розпізнавання мови

ABSTRACT

Bogodenko DO. *Models of speech signals and methods of their digital processing: bachelor's thesis: 171 Electronics. - Kyiv, 2022. - 95 p*

Thesis was completed on 87 sheets, it contains 1 appendix and a list of references to the sources used from 44 titles. The paper presents 22 figures.

Keywords: SPEECH SIGNALS, LANGUAGE RECOGNITION, RECOGNITION SYSTEMS.

Objective: Analysis of speech recognition technologies and study of their accuracy and effectiveness to improve performance.

Object of research: Language recognition systems and their functionality in use.

Subject of research: accuracy of speech recognition and functionality of systems.

The relevance of the topic of the bachelor's thesis is the study of speech recognition systems and their improvement for dissemination in society for a wide range of tasks

Research methods. The research used methods of theoretical synthesis to work with research literature and existing problems in the field of language recognition; a method of scientific generalization to prepare conclusions at the end of each section and at the end of the work to identify the factors of undergraduate research.

Scientific novelty: is to invent the latest solutions to improve the accuracy, efficiency and reliability of language recognition

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, СКОРОЧЕНЬ І ТЕРМІНІВ.

API – Application Programming Interface;

MFCC – Mel-frequency;

DCT – Discrete Cosine Transform;

DTW – Dynamic Time Wrapping;

NN – Neural Networks;

IDE – Integrated Development Environment;

CDDL – Common Development and Distribution License Computing;

CAP – система автоматичного регулювання;

D – звукова розбірливість;

S – складова розбірливість;

W – словна розбірливість;

I – фразова розбірливість;

A – формантна розбірливвість;

P(E) – залежність коефіцієнтів сприйняття формант від відносного рівня інтенсивності формант (від ефективного рівня відчуття фомант);

S(A) – залежність складової розбірливості від формантної;

W(S) – залежність словної розбірливості від складової;

k(f) – ваговий коефіцієнт, що характеризує ймовірність наявності формант промови в смузі частот;

LPC – Linear Predictive Coding;

АЧХ – амплитудно-частотна характеристика;

ПММ – прихована марківська модель;

DSP – цифрова обробка сигналів;

САРМ – система автоматичного розпізнавання мови.

ВСТУП

Ще до революційного винаходу Олександра Грема Белла інженери та вчені вивчали феномен мовленнєвої комунікації з ціллю створення ефективніших та ефективних систем спілкування між людьми. Починаючи з 1960-х років, цифрова обробка сигналів (DSP) відігравала центральну роль у вивченні мовлення, і сьогодні DSP є можпивістю реалізації плодів знань, отриманих протягом десятиліть досліджень. Одночасні досягнення в технології інтегральних схем і комп'ютерної архітектури створили технологічне середовище з практично безмежними можливостями для інновацій у додатках мовної комунікації. У цьому тексті висвітлюється центральна роль методів DSP у сучасних дослідженнях та застосуванні мовленнєвої комунікації. Представляється вичерпний огляд цифрової обробки мовлення, який варіюється від основної природи мовного сигналу, через різноманітні методи представлення мовлення в цифровій формі, до застосувань у голосовому спілкуванні та автоматичному синтезі та розпізнаванні мовлення. Широта цієї тематики не дозволяє обговорювати будь-який аспект обробки мовлення на велику глибину; отже, наша мета — надати корисне введення в широкий спектр важливих понять, які охоплюють область цифрової обробки мовлення. У зв'язку з цим, обробка мовних сигналів у сфері розвитку штучного інтелекту є одним із найважливіших завдань. Однак, перш ніж система повинна зрозуміти сенс висловлювання, їй необхідно визначити, якою мовою воно було вимовлено. Визначення розмовної мови у наші дні залишається актуальним завданням, оскільки досі немає методів, дозволяють однозначно вирішити це завдання. Основна складність її вирішення пов'язана, перш за все, з частотно-часовою варіативністю мовного сигналу (зміною його протяжності та частотних характеристик базових мовних одиниць - фонем), яка, як правило, обумовлена індивідуальними особливостями диктора,

довкіллям, емоційним забарвленням, акцентом і т.д. д. Для компенсації тимчасових змін давно розроблено та успішно застосовується система прихованих марківських моделей (ПММ). У той час, як обробка частотних параметрів сигналу менш однозначна, є безліч варіантів вирішення задачі, як на етапі обчислення параметричного вектора, так і на етапі класифікації. Процес частотно-часової обробки мовного сигналу отримав загальноприйняту назву – акустичне моделювання [1].

Поряд з цим, зараз, машинне навчання є областю наукових досліджень, що активно розвивається. Це як з можливістю швидше, простіше і дешевше обробляти дані, і з розвитком методів виявлення цих даних законів, якими протікають фізичні, біологічні, економічні та інші процеси. У деяких завданнях, коли такий закон визначити досить складно, використовують глибинне навчання (deep learning). Глибинне навчання розглядає методи моделювання високорівневих абстракцій даних за допомогою безлічі послідовних нелінійних трансформацій, які, як правило, представляються у вигляді штучних нейронних мереж. Головними недоліками традиційних методів визначення мови мовного сигналу є вимога точної розмітки навчальних даних та розробка мовних моделей, що є дуже трудомістким та дорогим процесом, а значить навчальна вибірка автоматично скорочується у розмірі. Застосування методів глибинного навчання дозволяє позбутися цих проблем. Щоб застосувати техніку цифрової обробки сигналів до завдань обробки мовлення, важливо розуміти основи процесу відтворення мови, і навіть цифрової обробки сигналів. У цьому розділі представлений огляд акустичної теорії відтворення мови і показано, як ця теорія призводить до безлічі способів уявлення мовного сигналу. Мовні сигнали складаються із послідовності звуків. Ці звуки і переходи між ними є символічним уявленням інформації. Розташування цих звуків регулюється правилами мови. Вивчення цих правил та їх значення у людському спілкуванні є

областю лінгвістики. Вивчення та класифікація звуків мови називається фонетикою. Голосовий шлях починається біля отвору між голосовими зв'язками або голосовою щілиною і закінчується біля губ, що складається з горлянки. Довжина голосу чоловіка = 17 см. Площа поперечного перерізу коливається від 0 до 20 см2. Носовий хід починається біля піднебіння а і закінчується у ніздрі. Підгортання система служить джерелом енергії для мови. Мова: хвиля, що випромінюється системою, коли повітря збуджене звуженням десь у голосовій доріжці. Звуки мови можна розділити на 3 різних класи в залежності від способу їхнього збудження. 1. Дзвінкий звук: створюється шляхом нагнітання повітря через голосову щілину з натягом голосових зв'язок, відрегульованим таким чином, щоб вони вібрували в коливаннях релаксації, тим самим виробляючи квазіперіодичні імпульси повітря, які збуджують голосовий слід. Приклад: У, д, ш, і, е. Фрикативні або глухі звуки: Утворюються шляхом утворення звуження в якійсь точці вокальної доріжки та нагнітання повітря через звуження із досить високою швидкістю, щоб викликати турбулентність. Це створює джерело шуму широкого спектра для збудження вокальної доріжки. Вибухові звуки: виникають у результаті повного закриття, наростання тиску за закриттям та його різкого відпускання. Питання людино-машинної взаємодії є одними з найважливіших під час створення нових комп'ютерів. Найбільш ефективними засобами взаємодії людини з машиною були б ті, що є природними для нього: через візуальні образи та мовлення. Створення мовних інтерфейсів могло б знайти застосування в системах різного призначення: голосове управління для людей з обмеженими можливостями, надійне управління бойовими машинами, що «розуміють» лише голос командира, автовідповідачі, що обробляють в автоматичному режимі сотні тисяч дзвінків на добу (наприклад, у системі продажу авіаквитків) тощо. При цьому, мовний інтерфейс повинен

включати два компоненти: систему автоматичного розпізнавання мови для прийому мовного сигналу і перетворення його в текст або команду, і систему синтезу мови, що виконує протилежну функцію - конвертацію повідомлення від машини в мову. Проте, незважаючи на стрімко зростаючі обчислювальні потужності, створення систем розпізнавання мови залишається надзвичайно складною проблемою. Це обумовлюється як її міждисциплінарним характером (необхідно мати знання у філології, лінгвістиці, цифровій обробці сигналів, акустиці, статистиці, розпізнаванні образів тощо), так і високою обчислювальною складністю розроблених алгоритмів. Останнє накладає суттєві обмеження на системи автоматичного розпізнавання мови – на обсяг словника, що обробляється, швидкість отримання відповіді та її точність. Не можна також не згадати про те, що можливості подальшого збільшення швидкодії ЕОМ за рахунок удосконалення інтегральної технології рано чи пізно будуть вичерпані, а зростаюча різниця між швидкодіями пам'яті та процесора лише посилює проблему. Існують області застосування систем автоматичного розпізнавання мови, де описані проблеми виявляються особливо гостро через жорстко обмежені обчислювальні ресурси, наприклад, на мобільних пристроях. Виробники мобільних телефонів і планшетів знайшли вихід у перенесенні ресурсомістких обчислень з пристроїв користувачів на сервери в хмарі, де, фактично, і розпізнавання. Додаток користувача тільки відправляє туди мовні запити і приймає відповіді, використовуючи підключення до інтернету. За цією схемою успішно працюють системи Siri від Apple та Google Voice Search від Google. Проте, для такої реалізації необхідні певні умови, наприклад, безперервний доступ до інтернету, які в ряді випадків недосяжні, і потрібно створити компактний і надійний самостійний пристрій, що експлуатує лише доступні «на місці» обчислювальні потужності. Описані проблеми виникають під час створення інтелектуальних механізмів як у

військовій сфері, так і у цивільної. Всі описані вище приклади поєднує необхідність створення компактного, надійного, самостійного та максимально швидкодіючого пристрою. Над розв'язанням зазначеної задачі працює безліч фахівців [2, 3]. Таким чином, пошук нових архітектурних рішень, що не базуються на архітектурі фон Неймана, є актуальною темою, особливо в її додатку до розв'язання задач штучного інтелекту, яких відноситься і розпізнавання мови. Одним з перспективних напрямів є розробка та дослідження асоціативних середовищ та побудови за допомогою них неоднорідних клітинних автоматів. Асоціативний доступ здійснюється, на противагу адресному, за вмістом інформації, а не за її адресою в середовищі, що запам'ятовує. Це дозволяє виконувати обробку інформації безпосередньо в середовищі логіки, а час асоціативного пошуку практично не залежить від ємності накопичувача. При цьому асоціативне осциляторне середовище складається з простих осередків, кожна з яких має свій закон функціонування, а разом ці осередки виробляють потокову обробку інформації. Успішне вирішення вищезгаданих завдань, а також прогрес обчислювальної техніки загалом дозволили звернутися до вирішення одного із завдань штучного інтелекту – автоматичного розпізнавання мови.

Мета роботи полягає у розробці методів розпізнавання мови в асоціативних середовищах та побудові системи розпізнавання у цих середовищах.

1. Історія та класифікація систем. Порівняльний аналіз існуючих методів розпізнавання людини по голосу.

1.1. Історія виникнення та розвитку систем розпізнавання мови.

Розпізнавання мови. як технологія, увійшла в суспільство порівняно недавно, з блискучими подіями запуску від

високотехнологічних гігантів провідних світових трендів. Історія виникнення систем розпізнавання мови дуже багата і різноманітна, яка починалась ще визначенням окремих слів, надалі ще більших словників і нарешті до швидких відповідей на запитання, як це робить наприклад Siri. Вперше засіб для розпізнавання мови з’явився ще в далекому 1952 році. Bell Laboratories розробили систему "Audrey", яка розпізнавала вимовлені людиною цифри. Це вірно, враховуючи складність мови, що інженери спочатку зосередились на цифрах. Та вже через десять років в 1962 році IBM випустила їх розроблену систему "Shoebox", яка розуміла 16 слів англійською. Лабораторії в США, Японії, Англії та СРСР розробили ще кілька апаратів, які розпізнавали окремі вимовлені звуки, розширивши технологію розпізнавання мови підтримкою чотирьох голосних і дев'яти приголосних звуків. Звучали вони не дуже добре, але ці перші спроби дали вражаючий старт, особливо якщо враховувати, наскільки примітивними були комп'ютери того часу. Завдяки новим підходам і технологіям словниковий запас подібних систем виріс з декількох сотень до декількох тисяч слів і мав потенціал розпізнавання необмеженої кількості слів. Однією з причин був новий статистичний метод, більше відомий як прихована марківських модель. Використовуючи шаблони для слів і звукові патерни, вона розглядала ймовірність того, що невідомі звуки могли бути словами. Ця база використовувалася іншими системами ще протягом двадцяти років.Тільки в 1997 році випустили перший у світі «безперервний розпізнавач мови», тобто більше не доводилося робити паузу між кожним словом, у вигляді програмного забезпечення Dragon's NaturallySpeaking. Цей винахід був здатний розуміти 100 слів за хвилину, він все ще використовується сьогодні в оновленій формі і користується попитом у лікарів. Проривом для технології розпізнавання мови стало машинне навчання, завдяки чому кількість помилок при розпізнаванні слів стала значно меншою. Google

об'єднав новітні технології з хмарними обчисленнями для обміну даними і це призвело до запуску додатка Google Voice Search для iPhone у 2008 році. Google ввів елементи персоналізації в свої результати пошуку за допомогою голосу, і використовував ці дані для розробки свого алгоритму Hummingbird, отримуючи набагато більше тонке розуміння використовуваної мови. Потім з'явилася Siri, яка так само, як і система Google Voice Search, покладається на хмарні обчислення. Siri використовує ті дані, які їй відомі про тебе, щоб згенерувати відповідь, яка випливає з контексту і відповідає на твій запит як деяка особистість. Розпізнавання мови перетворилося з інструмента у розвагу. Методи, які використовувались для здійснення цих проривів вперед, стали витонченішими в тій мірі, в якій вони тепер вільно симулюють принципи, засновані на схемах роботи людського мозку. Комп'ютери на базі хмарних обчислень увійшли в мільйони будинків і можуть контролюватися голосом, навіть пропонуючи інтерактивні відповіді на широкий спектр запитів. На сьогоднішній день технологія дійсно корисна для виконання повсякденних завдань. Запитайте у Siri, Cortana або Google, яка погода буде завтра, і вона надасть цілком виразний словесний звіт. Звичайно пристрій ще не повністю досконалий, але розпізнавання мови досягло зараз прийнятного рівня точності для більшості людей, причому всі основні платформи повідомляють про частоту помилок менше 5% [4].

Мова йде про нестаціонарний сигнал, тобто його характеристики змінюються у часі. Можна наочно зобразити ці зміни, побудувавши графіки модулів ДПФ для фрагментів (фреймів) мовного сигналу, що йдуть підряд. Зображення, що вийшло, називається спектрограмою (рисунок 1,3.б). На рисунках 1.1 і 1.2 видно, що найбільше енергії несуть частоти до 8 КГц. Тому при оцифровці мовного сигналу типовим

вибором частот дискретизації. Його характеристики змінюються у часі. Можна наочно зобразити ці зміни, побудувавши графіки модулів ДПФ для фрагментів (фреймів) мовного сигналу, що йдуть підряд. Зображення, що вийшло, називається спектрограмою (рисунок1. 3.б). На рисунках 1.2 і 1.3 видно, що найбільше енергії несуть частоти до 8 КГц. Тому при оцифровці мовного сигналу типовим вибором частоти дискредитації є 16 КГц. Як слова в писемному мовленні утворюються з кінцевого набору символів – алфавіту мови, так і усне мовлення при всій її

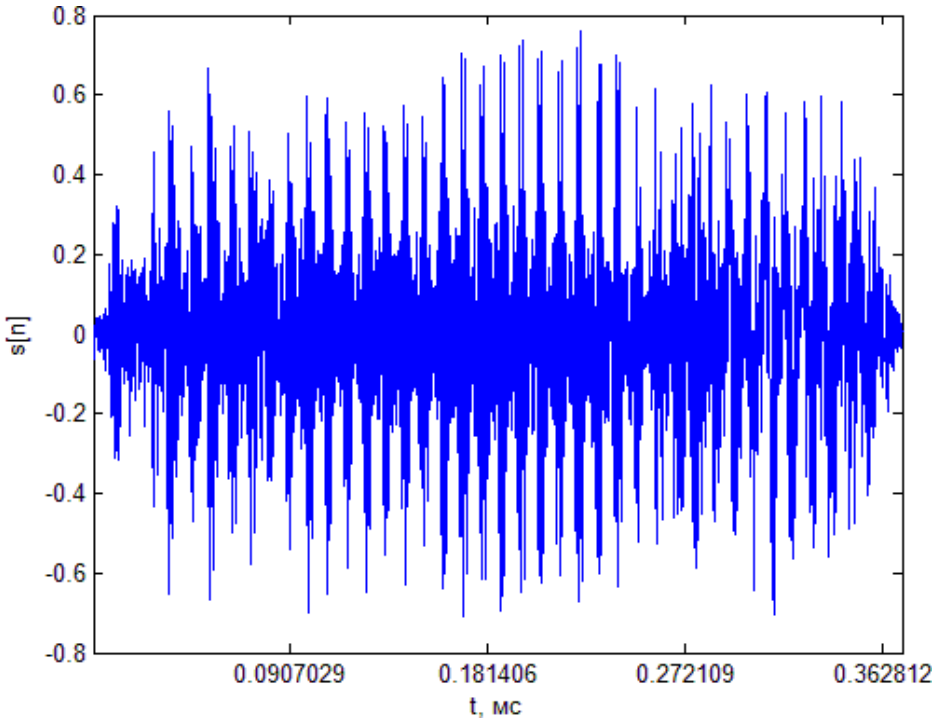


Рисунок 1.1. Ділянка промови з голосним звуком «а» [5]

варіативності включає обмежений набір звукових «літер». Мінімальною сенсорною одиницею мови є фонема.Фонетична система сучасної української мови, включаючи літературні норми й деякі діалектні особливості, налічує 38 основних фонем: 6 голосних і 32 приголосних; додатково визначають 13 приголосних фонем у периферійній підсистемі, але існують й інші погляди на фонемний склад української мови.

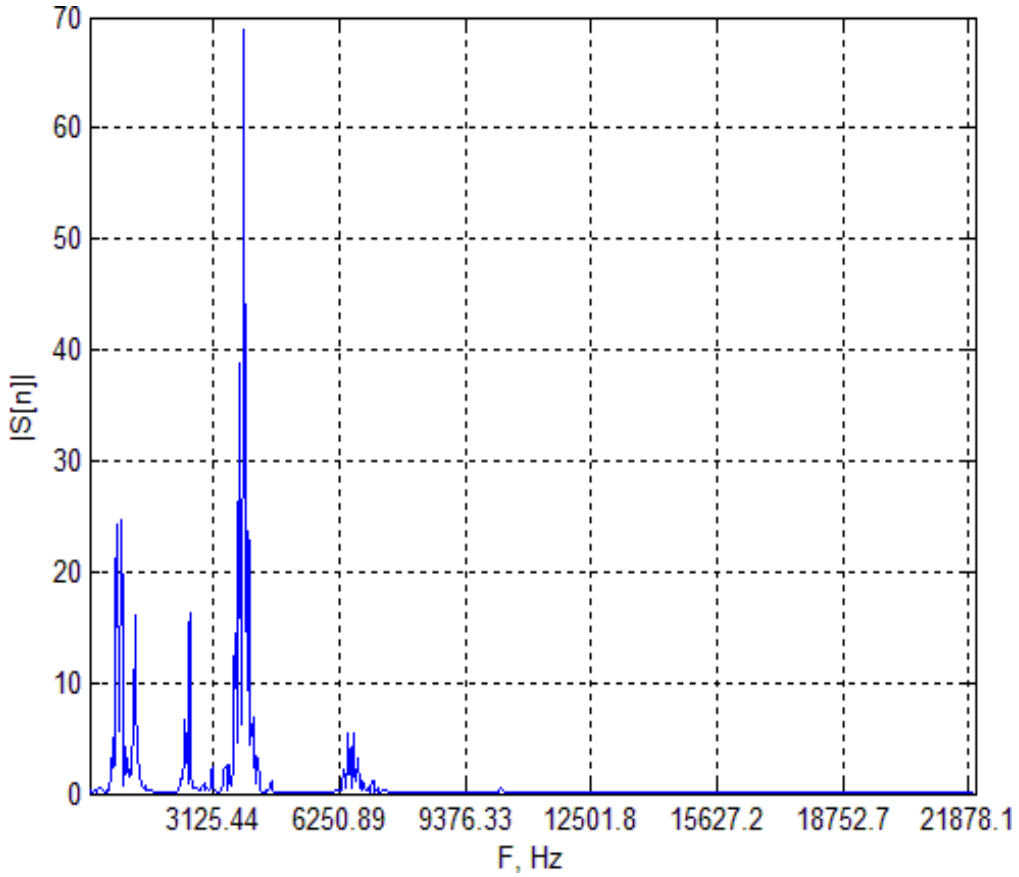
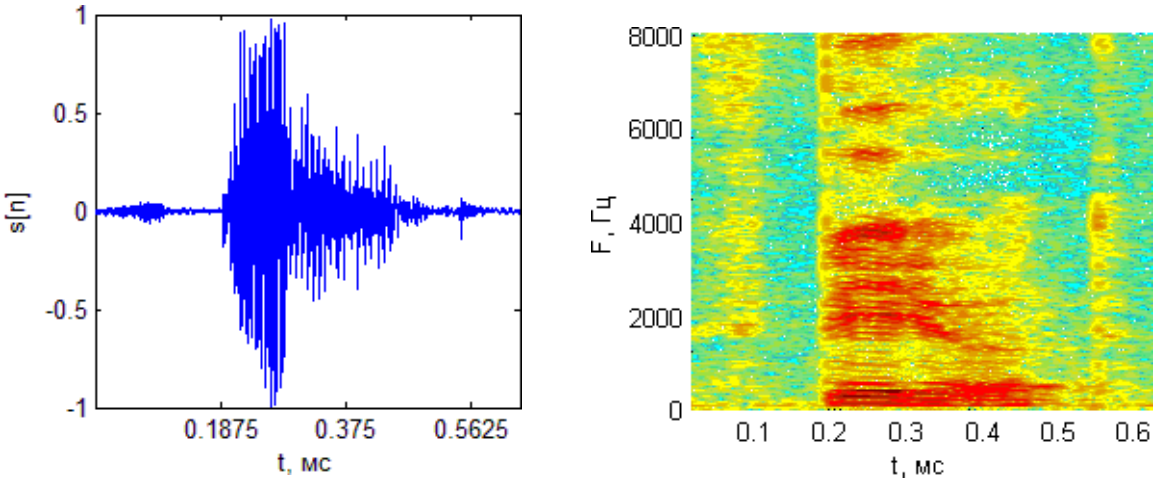


Рисунок 1.2. ДПФ для дільниці мовленого сигналу с голосним звуком «а» [6]



а) б)

Рисунок 1.3. Осциллограмма слова «Вперед» і його спектрограма [6].

1.2. Загальна структура системи автоматичного розпізнавання мовлення.

Більшість сигналів (мовних зокрема) мають аналогову природу, для обробки їх на цифрових комп'ютерах вони перетворюються на дискретні сигнали у вигляді аналого-цифрового перетворення (АЦП). За допомогою цієї процедури отримують набір відліків $s[n]$ – знятих у моменти $\Delta t \cdot n$ миттєвих значень безперервного сигналу, які вже позбавлені фізичної природи, а їхнє максимальне та мінімальне значення задається розрядністю АЦП. Наприклад, якщо розрядність АЦП дорівнює 2 байтам, то всі значення у відліках укладаються в проміжок 2^{16-1}, $2^{16-1} - 1$. При цьому найважливішим параметром перетворення є частота дискретизації, що визначає, скільки миттєвих значень безперервного сигналу (відліків) буде збережено за одну секунду . Частота дискретизації – величина, обернена до кроку дискретизації Δt. За теоремою Котельникова, з дискретного сигналу можна відновити без втрат тільки такий аналоговий сигнал, верхня частота $f_в$ спектра якого вдвічі менша за частоту дискретизації f_s.

$$fs > 2 \cdot fв \qquad (1.1)$$

Для опису та перетворення дискретних сигналів застосовні засоби цифрової обробки сигналів (ЦОС). Найважливішою процедурою ЦОС є дискретне перетворення Фур'є (ДПФ)

$$S\ m\ = \sum_{i=1}^{N} s[n] \cdot e^{-\frac{j2\pi nm}{N}},\ m = 1, \dots, N \qquad (1.2)$$

де N – кількість відліків, за якими будується ДПФ; j – уявна одиниця.

У роботі системи автоматичного розпізнавання мови (АРМ) вхідний мовленнєвий сигнал перетворюється в послідовність акустичних

векторів-ознак

$$\mathbf{Y}_{1:T} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T) \quad (1.3)$$

в результаті постпроцесингу. Потім декодер намагається відшукати послідовність мовленнєвих сегментів, заданих символами

$$\boldsymbol{w}_{1:L} = (w_1, w_2, ..., w_L)), \quad (1.4)$$

яка найбільш ймовірно відповідає ***Y,*** яка спостерігається:

$$\hat{\mathbf{w}} = \text{argmax}\, P(\mathbf{w} \mid \mathbf{Y}) \cong \text{argmax}\, p(\mathbf{Y} \mid \mathbf{w}) P(\mathbf{w}). \quad (1.5)$$

Еквівалентність правої частини виразу, що випливає із застосування правила Байєса, представляє базове формулювання генеративної моделі розпізнавання мовлення.

Акустична – $p(\boldsymbol{Y} \mid \boldsymbol{w})$ та лінгвістична – $P(\boldsymbol{w})$ – складові генеративної моделі, описуються кожна своїми стохастичними породжувальними граматиками.

Акустична модель кожного зі слів w формується в результаті композиції моделей базових мовленнєвих елементів, тобто фонем, які складають фонемну транскрипцію слова $\mathbf{q}^{(w)} = (q, q, ..., q)$. Для моделювання екстралінгвістичних явищ, властивих спонтанному мовленню, в алфавіт базових елементів, додатково до фонем і фонем-пауз, вводяться символи, які відображають неінформативні звуки. Загальноприйняті системи пофонемного розпізнавання оперують алфавітом фонем, контекстно-залежних або контекстно-незалежних, з яких будуються мовленнєві образи слів. Вже на послідовності слів накладаються обмеження шляхом введення лінгвістичної моделі на основі породжувальних граматик або статистичної моделі, враховуючи контексти слів. Загально прийняті системи розпізнавання оперують алфавітом фонем, контекстно-залежних або контекстно-незалежних, з яких будуються мовленнєві образи слів. Вже на послідовності слів накладаються обмеження шляхом введення лінгвістичної моделі на основі граматик або статистичної моделі,

враховуючи контексти слів.

На рис. 1.4 зображено загальну структуру системи автоматичного розпізнавання злитого мовлення, яка є спільною для ізольованих, клієнт-серверних і гібридних технологій. первинних векторів-ознак [7].

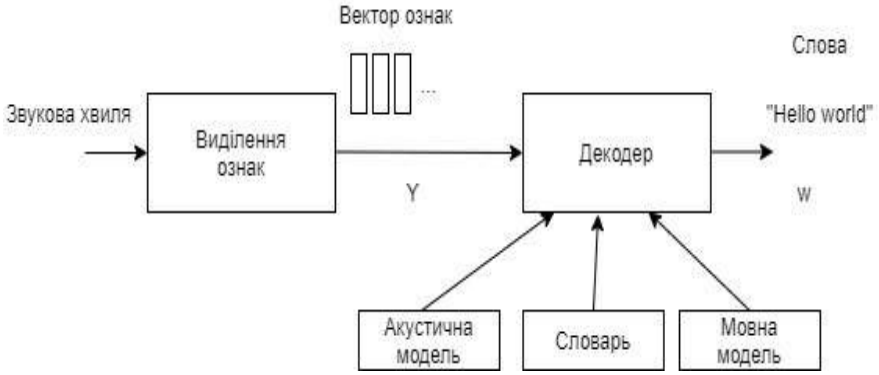


Рис.1.4. Система автоматичного розпізнавання злитого мовлення

При цьому використовується мілкепстральне перетворення з відніманням серед-нього значення. *Декодер* проводить порівняння вхідного мовленнєвого сегмента з гіпотезами еталонного сигналу допустимих послідовностей слів із *робочих словників*, застосовуючи деяку обережну стратегію відкидання малоперспективних гіпотез. Для цього використовуються дані з *акустичної* та *лінгвістичної моделей*. Послідовність слів, за якою генерується еталонний сигнал, найбільш схожий на вхідний сигнал, оголошується відповіддю розпізнавання. В ізольованих системах розпізнавання мовлення модулі реального часу та всі компоненти бази даних і знань знаходяться безпосередньо на портативному пристрої. Клієнт- серверні системи передбачають розміщення детектору голосової активності на пристрої, у той час як препроцесор може перебувати як на серверній, так і на клієнтській частині. Декодер із базами даних і знань перебуває на сервері. У гібридних системах розташування як модулів, так і компонент бази даних і знань варіюється.

1.3. Типи систем розпізнавання мови за послідовністю слів

Усі сучасні описи мовлення певною мірою є імовірнісними. Це означає, що між одиницями чи між словами немає певних пауз. В системах розпізнавання мови можна умовно виділити декілька класів, які послідовності слів вони мають можливість аналізувати окремі слова. Такі системи розраховані на то, що вимова кожного слова буже оточено тишою з обох сторін, тобто до і після слова повинна слідувати пауза. Зручно використовувати у ситуаціях, коли потрібно сказати якусь команду, яка складається з одного слова. В такій системі окремою задачою виявляється виділення пауз у мові між сусідіми входженнями, однак таку задачу все ж таки найлегше реалізувати навідміну від наступних. Йдеться про нестаціонарниий сигнал, тобто. його характеристики змінюються у часі. Можна наочно зобразити ці зміни, побудувавши графіки модулів ДПФ для фрагментів (фреймів) мовного сигналу, що йдуть підряд. Зображення, що вийшло, називається спектрограмою (рисунок 1.3.б). На рисунках 1.2 і 1.3 видно, що найбільше енергії несуть частоти до 8 КГц. Тому при оцифровці мовного сигналу типовим вибором частоти дискретизації є 16 КГц [8, 9].

1.3.1. Типи систем розпізнавання мови за розміром словника

Розмір лексики має важливе значення для системи розпізнавання мови та її роботи. Словниковий запас визначається як набір слів, з яких система може вибрати, тобто слова, на які він може посилатися. У випадках, коли варіантів небагато, розпізнавання очевидно простіше, ніж якщо словниковий запас великий. Так виділяють такі типи на основі об'єму словника:

-Малий словник (розпізнає від 1 до 100 слів або фраз);

-Середній словник (розпізнає від 101 до 1000 слів або фраз);

-Великий словник (розпізнає від 1001 до 10000 слів або фраз) [10].

1.3.2. Типи систем розпізнавання мови по залежності від диктора.

Кожна людина має свої особливості та властивості голосу, які будуть розглянуті далі в цій главі. Базуючись на цьому можна виокремити дві типи систем:

Дикторозалежні моделі.

Програмне забезпечення для розпізнавання мови, яке залежить від знання особливостей голосу мовця. Програмне забезпечення вивчає характеристики голосу мовця через голосове навчання. Цей тип системи повинен бути навчений конкретному користувачеві, перш ніж він зможе розпізнати сказане. Цей тип системи працює добре, якщо буде лише один користувач, який буде говорити з системою. Системи, що залежать від ораторів, як правило, здатні розпізнавати мовлення з різних контекстів (слів, фраз). Однак незалежні від оратора системи здатні розпізнавати мовлення у різних користувачів, обмежуючи контексти мови (слова та фрази) [11].

Дикторонезалежні моделі.

Програмне забезпечення, яке не потребує навчання, не залежить від ораторів. Ці системи використовуються для автоматизованих телефонних інтерфейсів. Ці системи можуть бути використані різними особами без навчання розпізнаванню мовленнєвих особливостей кожної людини [11].

1.3.3.Порівняння існуючих систем

Розглянемо найпопулярніші постачальники послуг з розпізнавання голосу на ринку.

Dragon Naturally Speaking (його ще називають Dragon for PC) є найкращим загальним програмним забезпеченням для розпізнавання

голосу. Його можна використовувати як в особистих, так і в офіційних цілях. Так Dragon Home допоможе вам у кількох повсякденних заходах, таких як диктування домашніх завдань, надсилання електронної пошти та навіть у веб-серфінгу. Dragon Professional Individual допомагає працюючим особам та малому бізнесу створювати та переписувати документи, вставляючи підпис або налаштовуючи словниковий запас. Dragon Legal Individual - це допомога в юридичних професіональних справах таких, як впорядкування юридичної документації [12, 13].

Google Cloud Speech API є лідером індустрії. Це програмне забезпечення розпізнає 120 мов, яке навчилось розпізнавати різних спікерів і самостійно визначати мову записи з декількох можливих. Google Cloud Speech API розшифровує аудіо з кол-центрів, перекладає, підтримує глобальних користувачів, самостійно розставляє знаки пунктуації та виконує команди. Він навіть використовує власний "мозок" для транскрипції попередньо записаного звуку. Це дивовижний інструмент. Також він підтримується всією підтримкою, якою Google так відомий [14].

Amazon Lex надає інструменти для вирішення складних завдань глибокого навчання, таких як розпізнавання мови і розуміння мови, за допомогою простого і повністю керованого сервісу. Amazon Lex інтегрований з сервісом AWS Lambda, який можна використовувати для простого доступу до більшості функцій і виконання функціонального коду на сервері з метою вилучення та оновлення даних. Після створення бот можна розгортати безпосередньо на платформі чату, мобільному клієнті або пристрої IoT. Можна також використовувати надані звіти для відстеження метрик свого бота. Amazon Lex надає безпечне і просте у використанні комплексне рішення, що дозволяє створювати, публікувати боти і відстежувати їх стан [15].

Усім відомий Siri- це найпопулярніший особистий помічник в Америці. Siri може дзвонити і відправляти повідомлення, коли ви за кермом, у вас, що може знадобитися, наприклад, запропонує відправити повідомлення колегам, якщо ви спізнюєтеся на зустріч. Siri допоможе працювати з пристроєм, не торкаючись до нього. Розумний помічник запам'ятовує ваші звички і передбачає, що вам може знадобитися протягом дня. Ви можете перевіряти факти, робити розрахунки або переводити фрази з однієї мови на іншу. З 95-відсотковою точністю Сірі перевершує всіх своїх гігантів з Кремнієвої долини. Точність та інтелект помічника і надалі будуть продовжувати вдосконалюватися.

Наступний лідер ринку Voice Finger використовується в цілях схожих до нашої розробленої програми. За допомогою Voice Finger ви зможете керувати комп'ютером лише голосом і не потрібно буде використовувати клавіатуру та мишу. Він підтримує команди Windows. За допомогою цього інструменту ви зможете виконувати завдання з нульовим контактом на комп’ютері [16].

1.3.4. Синтезатори мовлення

У той час як завдання розпізнавання мови дуже складна і вирішена лише частково, завдання синтезу мови набагато простіше (хоча і там є чимало проблем, що чекають свого рішення). У побуті нам доводиться стикатися з різними системами синтезу мови. Ось кілька прикладів. Технології синтезу мови застосовуються в метро при оголошенні зупинок.

Власники мобільних телефонів можуть спілкуватися з автоматичною сервісною службою для визначення залишку коштів на рахунку, перемикання тарифних планів, підключення або відключення послуг тощо. Сервісна служба спілкується голосом із застосуванням

технологій синтезу мови [17, 18].

Випущено чимало дитячих іграшок, що «говорять» людським голосом. У цих іграшках також застосовуються найпростіші синтезатори мови або цифрові магнітофони. Синтезатори мови застосовуються в різних голосових системах попередження, що встановлюються в автомобілях і літаках. Такі системи дозволяють привернути увагу людини до виникнення тієї чи іншої критичної ситуації, не відволікаючи його від процесу керування автомобілем, літаком чи іншим аналогічним засобом.

Також розроблено чимало комп'ютерних програм, здатних читати голосом вміст текстових файлів або текст, розташований на панелі програм. Ці системи можуть виявитися корисними тим, у кого ослаблений або повністю відсутній зір.

1.3.5.Моделі синтезу мови

Всі існуючі в даний час методи синтезу людської мови засновані на використанні двох моделей - моделі компілятивного синтезу і формантно-голосові моделі, як показано на рисунку 1.5.

Рисунок 1.5 – Моделі синтезаторів мовлення

Розглянемо детальніше кожну з цих моделей нижче:

Модель компілятивного синтезу передбачає синтез мови шляхом конкатенації (складання) записаних зразків окремих звуків, вимовлених диктором. При використанні цієї моделі складається база даних звукових фрагментів, з яких в подальшому буде синтезуватися мова. На перший погляд цей підхід не повинен викликати особливих труднощів. Модель компілятивного синтезу підходить, головним чином, тільки в найпростіших випадках, коли синтезатор повинен вимовляти відносно невеликий і заздалегідь відомий набір фраз. При цьому забезпечується досить висока якість мови. Втім, цей факт не дуже дивний, якщо згадати, що для синтезу використовується природна людська мова. Проте, на стику складених звукових фрагментів можливі інтонаційні викривлення і розриви, помітні на слух. Крім того, створення великої бази даних звукових фрагментів, що враховує всі особливості вимови фонем і алофонів з різними інтонаціями, являє собою складну і копітку роботу.

Формантно-голосова модель заснована на моделюванні мовного тракту людини. Ця модель може бути реалізована із застосуванням нейронних мереж і допускає самонавчання. На жаль, зважаючи на складність точного моделювання особливостей мовного тракту, а також з врахуванням інтонаційної модуляції мови формантно-голосова модель володіє відносно низькою точністю синтезованих звуків мови. Проте, сучасні програми синтезу мови, побудовані з використанням цієї моделі, синтезують цілком розбірливу мову і можуть застосовуватися в ряді випадків. Треба зауважити, що системи голосового попередження про виникнення аварійних ситуацій краще будувати з використанням моделі компілятивного синтезу, так як розбірливість мови в таких системах виходить на передній план [19].

Що ж стосується «побутових» синтезаторів мови, то в них можна з успіхом застосовувати і формантно-голосову модель.

1.3.6.Порівняльний аналіз синтезаторів мовлення.

Нижче наведено інформацію про декілька програмних реалізацій синтезаторів мови. Більшість таких синтезаторів розроблено для платформи Microsoft Windows і користується мовним програмним інтерфейсом Speech API, розробленим компанією Microsoft.

Синтезатор мови Google дозволяє озвучувати текст в додатках.

Наприклад, він використовується:

-в Google Play Книгах для читання вголос;

-в Google Перекладач для вимови слів;

-в TalkBack і інших сервісах спеціальних можливостей для озвучування тексту на екрані;

-в інших додатках, які можна знайти в Play Маркеті.

SpeechText – технологія, яка розроблена спеціально, для тих, у кого зайняті руки, кому цікаво знати вимова будь-яких слів. Цей додаток вимовляє будь-який введений вами текст, на будь-якому з доступних мов. Є можливість зберігати в вигляді звукового файлу на карту пам'яті те, що було написано.

Болтун – зручний додаток для пристроїв на Андроїд. Може озвучити текст SMS, повідомлення електронної пошти, статтю в браузері - будь-які тексти, які ви скопіювали в буфер обміну або введете прямо в додаток [16, 17].

Висновки до розділу

1. Досліджено історію виникнення та розвитку систем розпізнавання мови.

2. Показано, що за сучасними тенденціями завдання є дуже

актуальним.

3. Розглянуто основні ознаки, за якими можна класифікувати системи розпізнавання людської мови і те, як ці ознаки можуть впливати на роботу системи.

4. Наведений порівняльний аналіз декількох найбільш відомих та кращих існуючих систем. Окремо було досліджено синтезатори мови, а саме моделі синтезу мови та виконано порівняльний аналіз існуючих платформ.

Розділ 2. Структура мовлення та архітектура розпізнавання, загальні поняття про нейронну мережу

2.1.Структура мовлення

Мова є найбільш природною формою людського спілкування і тому реалізація інтерфейсу з урахуванням аналізу мовної інформації є перспективним напрямом розвитку інтелектуальних систем управління. Однією з актуальних невирішених завдань області інформаційно-вимірювальних систем є створення систем автоматичного розпізнавання інваріантних для диктора голосових сигналів. Її рішення дозволило б розширити коло користувачів таких систем та значно підвищити ефективність обміну інформацією у людино-машинних системах. Реалізація мовного інтерфейсу — дуже складне технічне завдання, вирішення якого перебуває в стику багатьох галузей науки. Тому при сприйнятті мови людина використовує механізми асоціативного аналізу, не тільки аналізуючи та порівнюючи почуті звуки, але збираючи фонеми в словесні образи, підбираючи найбільш підходящі слова не тільки за подібністю звучання, але й за інтонацією, емоційне забарвлення, контекст слів, словосполучень, речень навіть усього тексту. Тому людина здатна розпізнавати мову навіть за великої нестачі інформації. Наприклад, людина набагато вимогливіша до якості звуку при прослуховуванні

тексту іноземною мовою, який він погано знає, ніж при сприйнятті рідної мови. Ще час безпосередньо впроваджувати лінгвістичний інтерфейс у повсякденне життя кінцевого користувача, але поточний прогрес складно переоцінити. Програми та системи, що мають засоби лінгвістичного введення, набувають все більшого поширення, але, враховуючи всі їхні недоліки, варто розглянути перспективи розробки вузькоспеціалізованих систем, які мають чітке застосування [20].

2.2.Архітектура систем розпізнавання

Фаза зустрічі функцій має здатність створити образ форми голосових зв'язок. Тобто в цифровому сигналі виділяються області, які містять розмовну частину, виключаються проміжні та знаходять параметри мовлення. Основна мета цього етапу полягає в тому, щоб позбавить нас від використання необхідної несуттєвої інформації, фону, емоцій та іншого. Вектори функціональності полягають у генерації ділень у кадрі довжиною приблизно 25 мс і обчисленні 10 мс. Одна зі схем кодування для зразків і використання заснована на коефіцієнті Mel-Frequency Cepstral Coefficient (MFCC). Основа розуміння мови полягає в тому, що вони дуже відрізняються за формою голосового тракту, включаючи язик, а також за зубами та оком. Ця форма визначається тим, чим вона є. Почнемо з окремого у формі, слід дати уявлення про фонемну точність твору. Форма вокального тракту проявляється у смаку енергетичного спектру пивоварного терміну, і збірка MFCC точно представляє цей смак. Коефіцієнти кепстрал частоти Mel (MFCC) – це функція, яка амбітно використовується для автоматичного розпізнавання гучномовців і гучномовців. Його познайомили Девіс і Мермельштейн у 1980-х роках і в стані модернізму. До впровадження MFCC основними функціями автоматичного розпізнавання мови були коефіцієнти лінійного передбачення (LPC) і коефіцієнти лінійного прогнозування

(LPCC). Нижче наведені основні аспекти коефіцієнта частоти цього, важливо впроваджувати та впроваджувати спеціальні заходи для відновлення. Алгоритм відстеження та частотні коефіцієнти цієї частоти для даного кадру виглядає наступним чином: Мел-частотні кепстральні коефіцієнти (MFCC) - це функція, яка широко використовується в автоматичному розпізнаванні мови та динаміків. Вона були представлені Девісом і Мермельштейном у 1980-х роках і з тих пір є найсучаснішою. До впровадження MFCC, лінійні коефіцієнти прогнозування (LPC) та лінійні коефіцієнти прогнозування (LPCC) були головними функціями для автоматичного розпізнавання мови. Далі викладені основні аспекти мел-частотних кепстральних коефіцієнтів, а саме як їх реалізувати та чому_вони_є_особливими_для_розпізнавання [21].

Алгоритм знаходження мел-частотних кепстральних коефіцієнтів для заданого фрейму наступний:

1. Розкладаємо у ряд Фур’є.

Вихідний мовний сигнал запишемо в дискретному вигляді як:

$$x[n], 0 \leq n \leq N \quad (2.1)$$

Застосуємо до нього перетворення Фур’є, а саме можна використати швидке перетворення Фур’є (FFT реалізація):

$$Xa\ [k] \sum_{n=0}^{N-1} x[n]\mathrm{e}(-2\pi i/\mathrm{N})\mathrm{kn} \quad 0 \leq k \leq N \quad (2.2$$

Далі застосуємо віконну функцію Хеммінга для зглажування значень на межах фреймів:

$$H[k] = 0.54 - 0.46 \cos (2\pi k/N-1) \quad (2.3)$$

В результаті отримуємо наступний вектор:

$$X[k] = X[k] * H[k], 0 \leq k < N \quad (2.4)$$

Важливо розуміти, що після цього перетворення по осі X ми маємо частоту (hz) сигналу, а по осі Y - магнітуду (як спосіб піти від комплексних значень). Нижче на рисунку 2.1 наведені частота і магнітуда сигналу [22]:

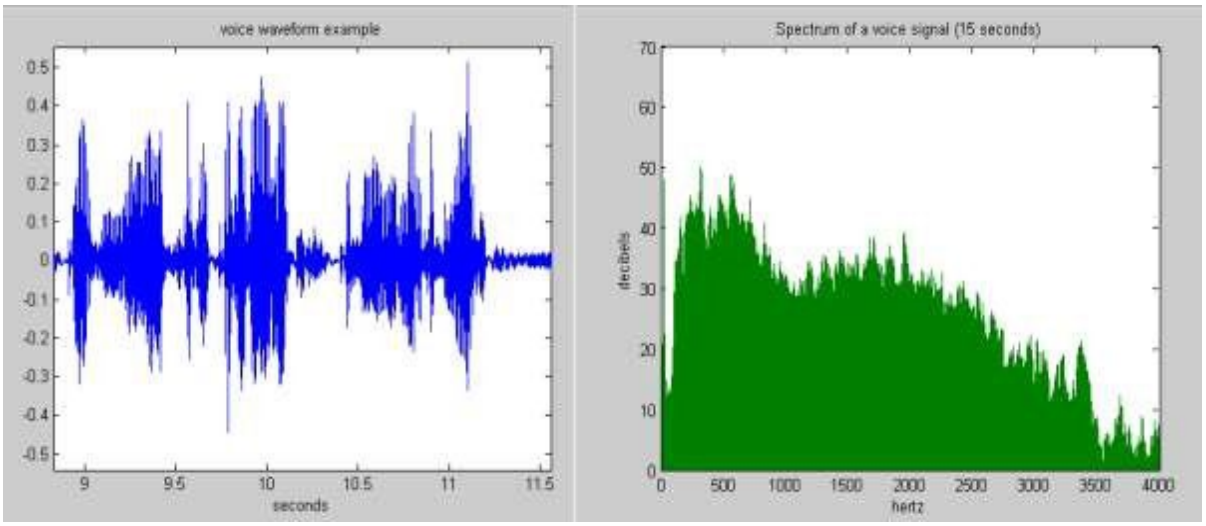


Рисунок 2.1 Частота та магнітуда сигналу

2. Розкладаємо спектр отриманий вище по mel-шкалі.

Мел – це одиниця висоти звуку, заснована на сприйнятті цього звуку нашими органами слуху. Як відомо, амплітудно-частотна характеристика людського вуха навіть віддалено не нагадує пряму, і амплітуда - не зовсім точна міра гучності звуку. Тому і ввели емпірично підібрані одиниці гучності, наприклад, фон. В першу чергу мел одиниця залежить від частоти звуку (а також від гучності і тембру), як показано на рисунку 2.2.

Аналогічно, сприймається людським слухом висота звуку не зовсім лінійно залежить від його частоти, як показано на рисунку 2.3.

Така залежність не претендує на велику точність, але зате описується наступною формулою:

$$M(f) = 1127 * \log(1 + f/700) \quad (2.5)$$

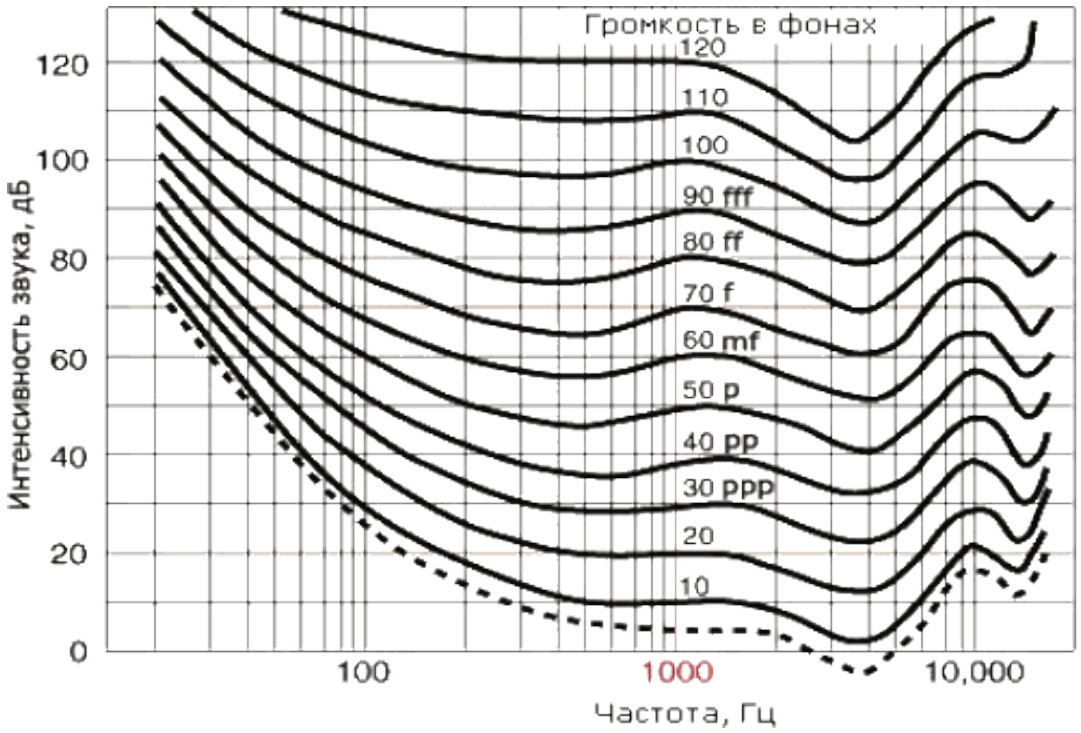


Рисунок 2.2 – АЧХ вуха людини [23]

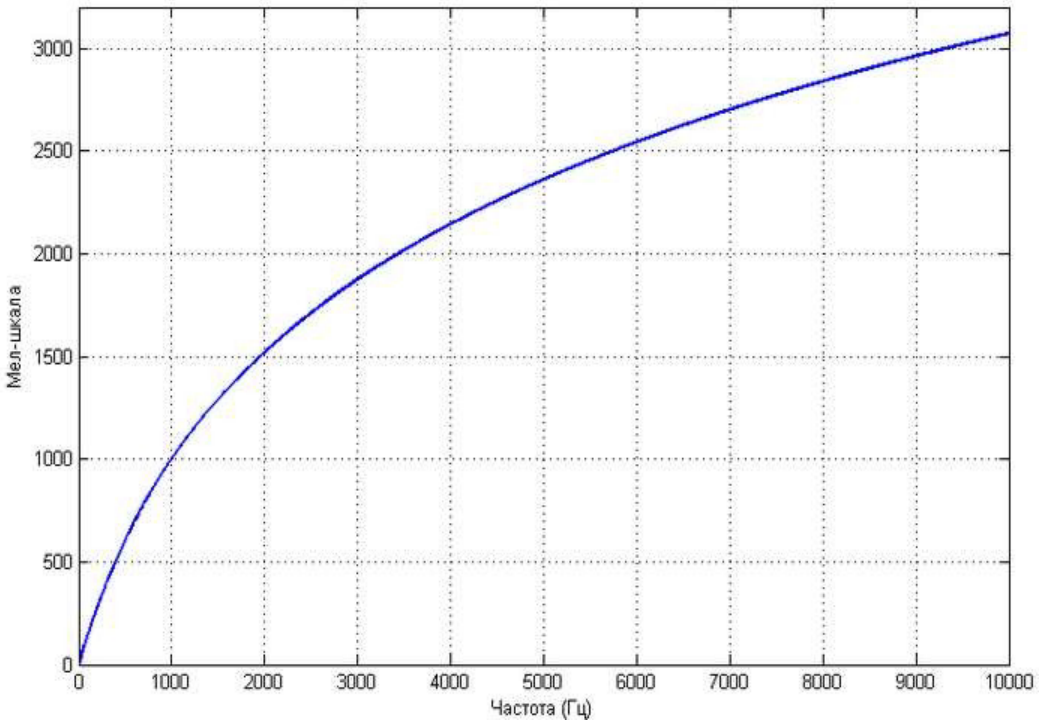


Рисунок 2.3 – Графік залежності mel-одиниць від частоти

Така залежність не претендує на велику точність, але зате виконується наступною формулою:

$$M(f) = 1127 * \log(1 + f/700) \quad (2.5)$$

Обернене перетворення має наступний вигляд:

$$M^{-1}(m) = 700 * * (\exp(m/1125) - 1) \quad (2.6)$$

Подібні одиниці виміру часто використовують при вирішенні задач розпізнавання, так як вони дозволяють наблизитися до механізмів людського сприйняття, яке поки що лідирує серед відомих систем розпізнавання мови [24].

Для розкладання спектру сигналу по mel-шкалі необхідно побудувати mel-фільтри. Мел-фільтр є трикутним віконною функцією, яка підсумовує енергію на своєму діапазоні частот і обчислює mel-коефіцієнти. Знаючи кількість mel-коефіцієнтів і діапазон частот можна побудувати набір наступних фільтрів, наприклад, як на рисунку 2.5

Чим більше порядковий номер мел-коефіцієнта, тим ширше основа фільтра. Це пов'язано з тим, що розбиття потрібному нам діапазону частот на оброблювані фільтрами діапазони відбувається на шкалі Мілов.

Нехай ми маємо фрейм з 256 елементами з частотою звуку 16000 Hz. Припустимо, що людська мова лежить в діапазоні від 300 до 8000 Hz. Кількість мел-коефіцієнтів, які ми хочемо знайти, рівна 10.

Застосувавши формулу (2.5) діапазон [300;8000] Hz зменшився до [401.25; 2834.99] Hz. Для побудови 10 трикутних фільтрів необхідно взяти 12 опорних точок: m[i] = [401.25, 622.50, 843.75, 1065.00, 1286.25, 1507.50,1728.74, 1949.99, 2171.24, 2392.49, 2613.74, 2834.99].

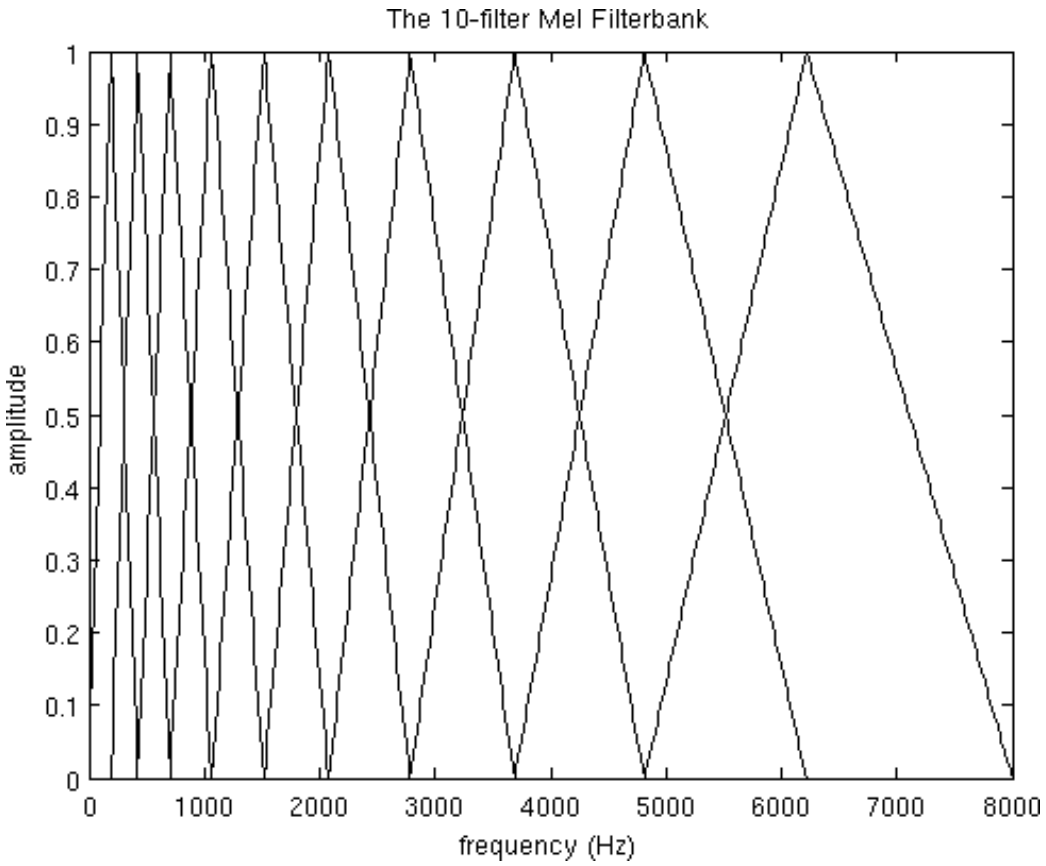


Рисунок 2.5 – Графік mel-фільтрів

Важливо, що точки на mel-шкалі мають рівномірний розподіл. Застосуємо формулу (6) для оберненого перетворення шкали в Hz. h[i] = [300, 517.33, 781.90, 1103.97, 1496.04, 1973.32, 2554.33, 3261.62, 4122.63, 5170.76, 6446.70, 8000]

Бачимо, що шкала поступово розтягується, вирівнюючи тим самим динаміку зростання "значущості" на низьких і високих частотах.

Накладемо отриману шкалу на спектр нашого фрейму. По осі X у нас знаходиться частота. Довжина спектра 256 елементів, при цьому в ньому вміщується 16000 Hz. Обрахувавши пропорцію отримуємо наступну формулу:

$$f(i) = floor\left((frameSize + 1) * (\mathrm{h(i)/sampleRate})\right) \qquad (2.7)$$

В нашому прикладі отримуємо такий результат: f(i) = 4, 8, 12, 17, 23, 31, 40, 52, 66, 82, 103, 128.

Знаючи опорні точки на осі X спектру, легко побудувати необхідні фільтри за такою формулою:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases}, \quad (2.8)$$

1.Фільтруємо та логарифмуємо енергію спектра.

Застосування фільтру полягає в попарном перемножуванні його значень зі значеннями спектра. Результатом цієї операції є мел-коефіцієнт. Оскільки фільтрів у нас M, коефіцієнтів буде стільки ж.

$$S[m] = \log(\sum_{k=0}^{N-1} |X[k]|^2 * H_m[k]) , 0 \le m \le M \quad (2.9)$$

Однак, нам потрібно застосувати мел-фільтри не до значень спектра, а до його енергії. Після чого прологарифмувати отримані результати. Вважається, що таким чином знижується чутливість коефіцієнтів до шумів.

2. Застосовуємо дискретне косинусе перетворення.

За допомогою дискретного косинусного перетворення (DCT) можна отримати "кепстральні" коефіцієнти.

Треба трохи розповісти і про друге слово в назві - кепстра. Мова являє собою акустичну хвилю, яка випромінюється системою органів: легкими, бронхами і трахеєю, а потім перетворюється в голосовому тракті. Якщо припустити, що джерела збудження і форма голосового тракту відносно незалежні, мовний апарат людини можна представити у вигляді сукупності генераторів тональні сигнали і шумів, а також фільтрів:

1. Генератор імпульсної послідовності (тонів)
2. Генератор випадкових чисел (шумів)
3. Коефіцієнти цифрового фільтра (параметри голосового

тракту)

4. Нестаціонарний цифровий фільтр

Основна задача дискретного косинусного перетворення – це "стиснення" отриманих результатів, причому з підвищенням значущості перших коефіцієнтів і зменшення значущості останніх. DCT рахуємо за такою формулою:

$$m{+}C[l] = \sum_{m=0}^{M-1} \quad \mathrm{S[m]}*\cos(\pi * \mathrm{l} * \frac{m+0.5}{M}), ,0 \leq l \leq \mathrm{M} \qquad (2.10)$$

У результаті для кожного фрейму ми отримаємо набір MFCC коефіцієнтів розміру M. Надалі їх можна буде використати у якості вхідних даних для акустичної моделі. Відповідно до структури мовлення, для розпізнавання мови використовується три моделі:

1. Акустична модель - це функція, що приймає на вхід невелику ділянку акустичного сигналу (кадр або фрейм) і видає розподіл ймовірностей різних фонем на цьому кадрі. Таким чином, акустична модель дає нам можливість по звуку відновити, що було вимовлено - з тим або іншим ступенем впевненості. Фонема - елементарна одиниця людської мови

2. Мовна модель використовується для обмеження пошуку слів. Вона визначає, яке слово могло б слідувати раніше розпізнаним словам (відповідність є послідовним процесом) і допомагає значно обмежити процес узгодження шляхом вилучення слів, які не вірогідні. Найпоширеніші мовні моделі - це n-грамові мовні моделі - вони містять статистику послідовностей слів – і моделі кінцевих мов - вони визначають мовленнєві послідовності методом автоматичного обмеження стану, іноді з вагами. Щоб досягти хорошого показника точності, мовна модель повинна бути дуже успішною в обмеженні простору пошуку. Це означає, що слід дуже добре передбачити наступне слово. Мовна модель, як правило, обмежує словниковий запас, який вважається словами, який він містить. Це проблема розпізнавання імен. Щоб вирішити це, мовна

модель може містити менші фрагменти, як підслова або навіть фонеми. Обмеження шуканого фрагмента в цьому випадку зазвичай гірше, а відповідні точності розпізнавання нижчі, ніж у мовній моделі на основі слів [25].

3. В ході роботи системи автоматичного розпізнавання мови завдання розпізнавання зводиться до визначення найбільш вірогідною послідовності слів, відповідних змісту мовного сигналу. Найбільш ймовірний кандидат повинен визначатися з урахуванням як акустичної, так і лінгвістичної інформації. Це означає, що необхідно проводити ефективний пошук серед можливих кандидатів з урахуванням різної ймовірнісної інформації. При розпізнаванні суцільної мови число таких кандидатів величезне, і навіть використання найпростіших моделей призводить до серйозних проблем, пов'язаних з швидкодією і пам'яттю систем. Як результат, це завдання виноситься в окремий модуль системи автоматичного розпізнавання мови, званий декодером. Декодер повинен визначати найбільш граматично ймовірну гіпотезу для невідомого висловлювання - тобто визначати найбільш ймовірний шлях по мережі розпізнавання, що складається з моделей слів (які, в свою чергу, формуються з моделей окремих фонів). Правдоподібність (likelihood) гіпотези визначається двома факторами, а саме можливостями послідовності фонів, що приписуються акустичної моделлю, і можливостями проходження слів друг за другом, обумовленими моделлю мови [26].

Ці три об'єднання об'єднані разом в двигун для розпізнавання мови.

2.3 Методи та алгоритми розпізнавання мови

Сучасні методи та алгоритми розпізнавання мовлення можна поділити на наступні класи:

Динамічне програмування - алгоритм динамічної

трансформації часової шкали (Dynamic Time Wrapping - DTW).

Приховані Марківські моделі (Hidden Markov Models - HMM).

Нейронні мережі (Neural Networks - NN).

У багатьох статтях, які відносяться до розпізнавання мови довгий час базовим підходом вважалися приховані марківскі моделі, проте останнім часом активно розвиваються методи, засновані на застосуванні нейронних мереж. Далі наводиться більш детальний огляд зазначених методів [26, 27].

2.3.1 Динамічне програмування

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 [28].

Існують 2 особливості застосування алгоритму:

1.Пряме порівняння числових форм сигналів. В цьому випадку, для кожної числової послідовності створюється нова послідовність, розміри якої значно менше. Алгоритм має справу з цими послідовностями. Числова послідовність може мати кілька тисяч числових значень, в той час як підпослідовність може мати кілька сотень значень. Зменшення кількості числових значень може бути виконано шляхом їх видалення між кутовими точками. Цей процес скорочення довжини числової послідовності не повинен змінювати свого уявлення. Безсумнівно, процес призводить до зменшення точності розпізнавання. Однак, беручи до уваги збільшення швидкості, точність, по суті, підвищується за рахунок збільшення слів в словнику [29].

2. Подання сигналів спектрограм і застосування алгоритму DTW

для порівняння двох спектрограм. Метод полягає в поділі цифрового сигналу на деяку кількість інтервалів, які будуть перекриватися. Для кожного імпульсу, інтервали дійсних чисел (звукових частот), буде розраховуватись ШПФ, і буде зберігатися в матриці звуковий спектрограми. Параметри будуть однаковими для всіх обчислювальних операцій: довжин імпульсу, довжини перетворення Фур'є, довжини перекриття для двох послідовних імпульсів. Перетворення Фур'є є симетрично пов'язаних з центром, а комплексні число з одного боку пов'язані з числами з іншого боку. У зв'язку з цим, тільки значення з першої частини симетрії можна зберегти, таким чином, спектрограмма представлятиме матрицю комплексних чисел, кількість ліній в такій матриці є рівною половині довжини перетворення Фур'є, а кількість стовпців буде визначатися в залежності від довжини звуку. DTW буде застосовуватися на матриці дійсних чисел в результаті сполучення спектрограми значень, така матриця називається матрицею енергії [30].

1. DTW алгоритми є дуже корисними для розпізнавання окремих слів в обмеженому словнику. Для розпізнавання швидкого мовлення використовуються приховані моделі Маркова. Використання динамічного програмування забезпечує поліміальную складність алгоритму: O (n2v), де n - довжина послідовності, а v кількість слів у словнику.

2. DWT мають кілька слабких сторін. По-перше, O (n2v) складність не задовольняє великим словників, які збільшують успішність процесу розпізнавання. По-друге, важко вирахувати два елементи в двох різних послідовностях, якщо взяти до уваги, що існує безліч каналів з різними характеристиками. Проте, DTW залишається простим в реалізації алгоритмом, відкритим для поліпшень і відповідним для додатків, яким потрібна проста розпізнавання слів: телефони, автомобільні комп'ютери, системи безпеки і т.д.

2.3.2 Прихована марківська модель

Прихована марківська модель (ПММ) представляє з себе систему, яка імітує випадковий процес, еволюція якого залежить лише від поточного стану моделі і не залежить від попередньої історії. У такому випадку виникає задача з наближенням значень прихованих (невідомих) моделей із заданою послідовністю спостереження значень.

Таким чином, марківський процес може зобразити як на рисунку 2.6 [31]:

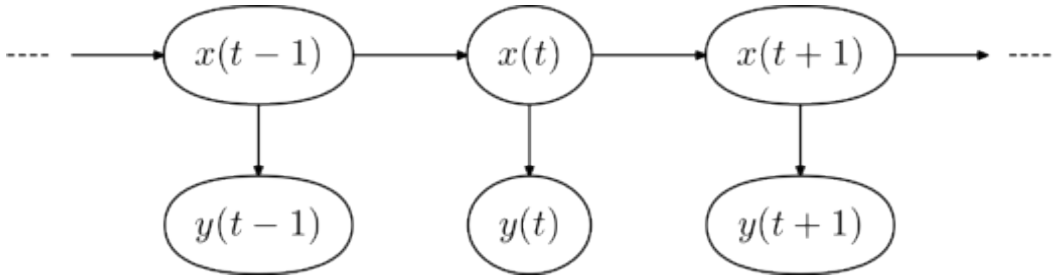


Рисунок 2.6. Марківський процес

Тут $x_t \in \{1, \dots, N\}$ - прихований стан моделі в момент часу t, $y_t \in R^d$-змінна, яка спостерігається в момент часу t. Значення y_t залежить тільки від x_t, а x_t тільки від стану в момент часу t - 1, тобто від x_{t-1}. Для подальшого задання моделі потрібно визначити матрицю ймовірностей переходів:

$$a \in R^{N\times N} | a_{i,j} = p(x_t = j | x_{t-1} = i), \qquad (2.11)$$

розподілу спостережуваних змінних для кожного стану $p(y_t|x_t)$ і розподіл ймовірностей початкових станів $p(x_1)$. Зазвичай розподіл:

$$p(y_t|x_t) = b_t\ , \qquad (2.12)$$

роблять нормальним:

$$b_t(y) = \mathcal{N}(y; \mu_t, \sum_t), \qquad (2.13.)$$

Форма вхідної звукової хвилі від мікрофона перетворюється в послідовність акустичних векторів фіксованого розміру $Y = \{y_1,\ y_2,\ ...\ y_T\}$, в процесі, який називається виділенням ознак. Декодер потім намагається знайти послідовність слів $w = \{w_1,\ w_2,\ ...\ w_L\}$, якими з найбільшою ймовірністю згенерована Y, тобто представимо наступним чином

$$\hat{w} = \arg\max\{P(w|Y)\}, \tag{2.14}$$

2.3.3 Нейронні мережі

Особливе місце в задачі розпізнавання мови займають методи, засновані на нейромережевих технологіях. У цих методах результат розпізнавання є продуктом функціонування нейронної мережі певного виду і топології. Нейронні мережі являють собою безліч пов'язаних між собою елементарних процесорів (нейроподібних елементів), кожний з яких виконує відносно прості функції [27, 28, 31].

Прототипом нейрона є біологічна нервова клітина. Нейрон складається з тіла клітини, або соми, і двох типів зовнішніх деревоподібних гілок: аксона і дендритів. Тіло клітини включає ядро, яке містить інформацію про спадкові властивості, і плазму, що володіє молекулярними засобами для виробництва і передачі елементів нейрона необхідних йому матеріалів. Нейрон отримує сигнали (імпульси) від інших нейронів через дендрити і передає сигнали, згенеровані тілом клітки, вздовж аксона, що наприкінці розгалужується на волокна, на закінченнях яких знаходяться синапси. Математична модель нейрона описується наступним співвідношенням:

$$y = f(s), s = \sum_{i=1}^{n} x_i w_i + b, \tag{2.14}$$

де wi – вага синапса; b – значення зміщення; s – вхідний сигнал;

y – вихідний сигнал нейрона; n – число входів нейрона

Структурна схема нейрона подана на рисунку 2.7:

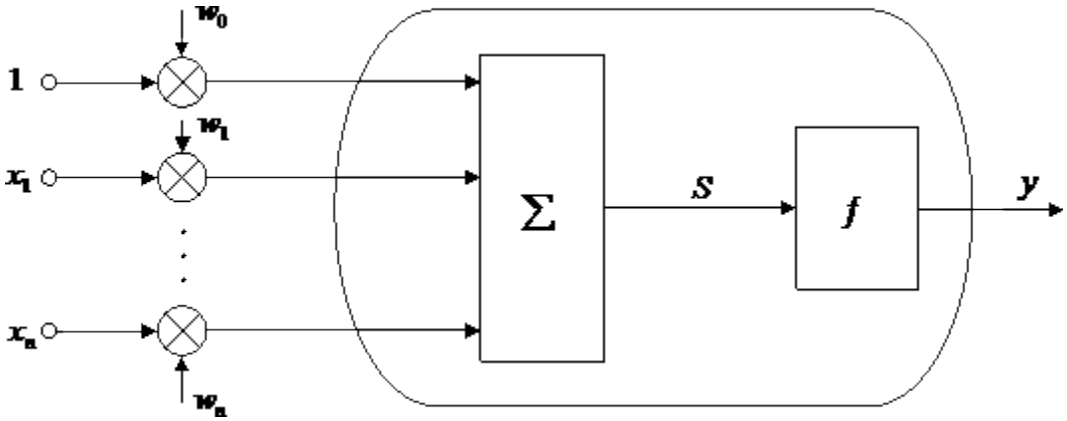


Рисунок 2.7 – Структурна схема нейрона

де x_1, x_2, ..., x_n - вхідний сигнал нейрона; w_1, w_2, ..., w_n - набір вагових коефіцієнтів; f (s) - функція активації; y - вихідний сигнал.

Як видно, нейроподібні елемент виконує нескладні операції зваженого підсумовування, обробляючи результат нелінійним пороговим перетворенням. Особливість нейросетевого підходу полягає в тому, що структура з простих однорідних елементів дозволяє вирішувати нетривіальні завдання завдяки складній організації зв'язків між елементами. Структура зв'язків визначає функціональні властивості мережі в цілому.

Процес функціонування мережі залежить від величин синаптичних зв'язків. Задавши певну структуру мережі (на етапі проектування), знаходять оптимальні значення вагових коефіцієнтів w_1, w_2, ..., w_n і зсувів b всіх нейронів. Цей етап називають навчанням нейронної мережі.

Вирішити поставлене завдання розпізнавання мови за допомогою нейронної мережі заданої структури - означає шляхом навчання за вибіркою, заданої k парами значень вхідних і вихідних векторів (X^i, Y^i), i=1,…,k, знайти таку конфігурацію, щоб забезпечити найбільш оптимальне в певному сенсі її функціонування. Розрізняють два підходи до навчання нейронної мережі: навчання з учителем і навчання без вчителя (самонавчання). При навчанні з

учителем на вхід мережі подається один з векторів X^i навчальної вибірки, вихід мережі Y порівнюється з виходом Y^i навчальної вибірки, і при необхідності робляться поправки в вагові коефіцієнти і зміщення нейронів мережі. В алгоритмах навчання без учителя підстроювання ваг синапсів проводиться на підставі інформації про стан нейронів і вже наявних вагових коефіцієнтів по одному з правил навчання. Процес повторюється, поки вихідні значення мережі не стабілізуються із заданою надійністю [33].

Нейрони можуть групуватися в мережеву структуру по-різному. Функціональні особливості нейронів і спосіб їх об'єднання в мережеву структуру визначають особливості нейромережі. Для вирішення задач ідентифікації та управління найбільш адекватними є багатошарові нейронні мережі (БНС) прямої дії, або багатошарові персептрони. При проектуванні БНС нейрони об'єднують в шари, кожен з яких обробляє вектор сигналів від попереднього шару. Мінімальною реалізацією є двошаровий нейронна мережа, що складається з вхідного (розподільного), проміжного (прихованого) і вихідного шару, як на рисунку 2.8

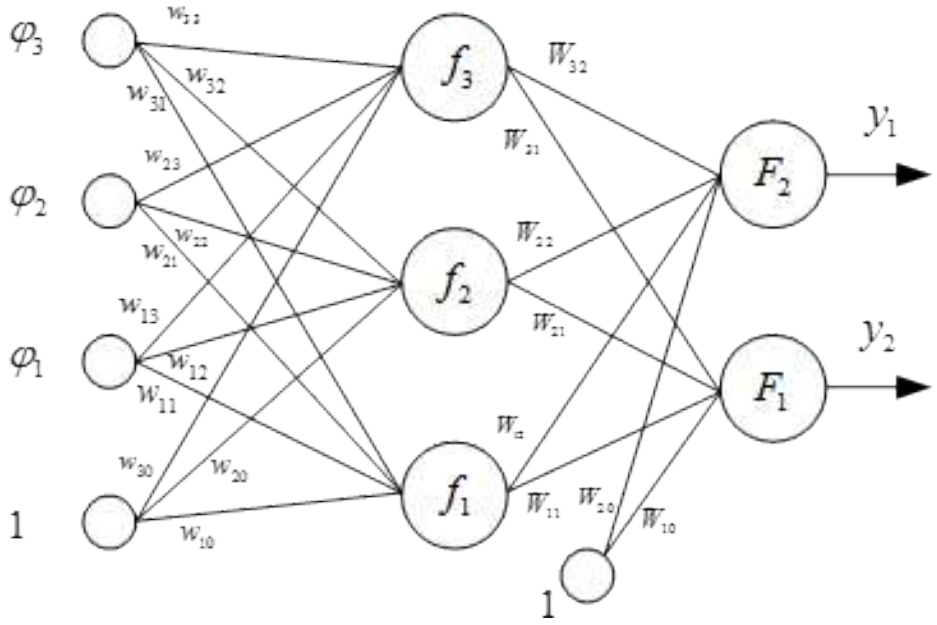


Рисунок 2.8 – Структурна схема двошарової нейронної мережі

Реалізація моделі двошарової нейронної мережі прямого дії має наступний математичний вигляд:

$$y(\theta) = \mathrm{F_i}(\sum_{j=1}^{\mathrm{nh}} \mathrm{Wijfj}\ (\sum_{j=1}^{nh} \mathrm{wij}\varphi\mathrm{j}\ +\ \mathrm{wj0}) + \mathrm{W_{j0}} \qquad (2.15)$$

n_φ – розмірність вектора входів φ нейронної мережі; θ – вектор параметрів, що настроюються нейронної мережі, що включає вагові коефіцієнти і нейронні зміщення (w_{ij}, W_{ij}); $f_j\ (x)$ – активаційна функція нейронів прихованого шару; $F_i\ (x)$ – активаційна функція нейронів вихідного шару.

Найважливішою відмінною рисою нейромережевого методу є можливість паралельної обробки. Дана особливість при великій кількості міжнейронних зв'язків дає можливість досягти значного прискорення процесу обробки даних. У багатьох випадках з'являється можливість обробки мовних сигналів в реальному часі. Ще один важливий плюс в нейромережевому методі - це узагальнення отриманих знань. Нейронна мережа має ті якості, які властиві для так званого штучного інтелекту.

В результаті процесу навчання нейронна мережа здатна виявляти складні залежності між вхідними та вихідними даними. При узагальненні інформації мережа дозволить повернути вірний результат на підставі неповних або перекручених даних. При великій кількості з'єднань між нейронами мережа набуває стійкість до помилок, які виникають на деяких лініях. Роботу пошкоджених зв'язків беруть на себе справні лінії, в результаті чого робота мережі не зазнає істотних змін.

Процедура навчання мережі є досить трудомістким в обчислювальному плані процесом, вимагає великого розміру навчальної вибірки. Крім того, процедура навчання не у всіх випадках гарантує отримання результату, багато робіт присвячено проблемам навчання нейронних мереж. Незважаючи на ці недоліки даного методу, він є одним з найчастіше використовуваних для розпізнавання мови, володіє наступними перевагами перед іншими методами: є нелінійним, досить стійкий до зашумлення мовного сигналу, легко піддається розпаралеливанню обчислень.

Розглянемо «класичний варіант» багатошарової мережі, де синаптичні зв'язки можуть визначатися будь-якими дійсними числами, а вихід нейрона - дійсними числами з інтервалу від 0 до 1. Як активаціонну функцію використовуємо сигмоїд, число шарів – довільне.

1.Визначаємо M матриць вагових коефіцієнтів W розміром NxN,

де M – число шарів,

N – число нейронів в одному шарі

$W_{i,j,k}$ буде позначати вагу j-го входу k-го нейрона в i-м шарі. Ініціалізуємо матриці деякими малими випадковими (не однаковими) значеннями.

2. Подаємо на входи мережі певні значення X, для яких відомі правильні значення виходів мережі Y*.
3. Обчислюємо значення виходів мережі для поточного стану матриць

W. Тобто для вхідного вектора X обчислюється вихідний вектор Y. Для цього необхідно послідовно обчислити вихід для кожного шару мережі з першого по останній. Для i-го шару в векторному вигляді це можна записати так:

$O_i = F(XW_i)$, якщо i – перший шар;

$O_i = F(O_{i-1}W_i)$, якщо i – не перший шар,

O_i – вектор виходу i-го шару; F – активаційна функція;

X – вектор входів;

O_{i-1} – вектор виходу (i-1)-го шару;

W_i – матриця вагових коефіцієнтів i-го шару.

1. Обчислюємо вектор

$$\Delta Y = Y - Y^* \qquad (2.16)$$

2. Якщо ΔY менше заданої похибки, переходимо до кроку 9.

3. Для шару з номером M (тобто в останньому шарі) виробляємо наступні операції:

3.1. Для всіх нейронів в шарі з номера 1 по N робимо наступні операції:

3.1.1. Для всіх ваг нейрона з номера 1 по N робимо наступні операції:

3.1.1.1Розраховуємо вектор

$$\delta M = X(1 - X)\Delta Y \quad (2.17)$$

3.1.1.2 Розраховуємо величину

$$\Delta W_{i,j,k} = \eta \delta i, kOi - 1, j \quad (2.18)$$

де η – коефіцієнт швидкості навчання (від 0.01 до 1.0).

1.1.1.1 Коригуємо величину вагового коефіцієнта, додаючи до $W_{M,j,k}$ величину $\Delta W_{M,j,k}$.

1. Для шарів з номером M-1 по перший послідовно проводимо такі операції:

1.1. Для всіх нейронів в шарі з номера 1 по N робимо наступні операції:

1.1.1. Для всіх ваг нейрона з номера 1 по N робимо наступні операції:

1.1.1.1. Розраховуємо вектор

$$\delta_i = O_{i+1}(1 - O_{i+1})[\sum \delta_{i+1,j} W_{i+1,j,k}] \quad (2.19)$$

Описаний алгоритм застосовується достатню кількість разів, щоб всі варіанти вихідних значень могли правильно виходити при завданні довільних значень входу із заданою вірогідністю помилки.

З усього вищесказаного можна зробити висновок про ефективність застосування нейромережевого методу до розглянутої задачі розпізнавання мови. Використання нейромережевих методів для розпізнавання мови дозволяє істотно підвищити точність системи розпізнавання мовного сигналу і збільшити її швидкодію.

Висновки до розділу

1.Детально розглянуто крок за кроком архітектуру розпізнавання мови. На першому етапі виділення ознак було детально розписано алгоритм знаходження мел-частотних кепстральних коефіцієнтів (MFCC) для заданого фрейму.

2. Наступний етап декодера включає в себе наведені три моделі: акустична, мовна та словник, які об'єднані разом. Детально оглянуто сучасні методи та алгоритми розпізнавання мовлення (Dynamic Time Wrapping, Приховані Марківські моделі, Нейронні мережі). Особливу увагу було приділено методу, заснованого на нейронні мережах, а саме детально розглянуто процедуру навчання «класичного варіанту» багатошарової мережі.

Глава 3. Системи і перетворення сигналів

У цьому розділі описані деякі типи систем і розглядаються дві основні завдання: перше – уявлення (моделювання) аналогових систем у цифровій формі, друге – оцінка параметрів дискретних систем.

Загальні питання моделювання сигналів та систем викладені в дипломі.

3.1 Загальні відомості про системи

Під час розгляду сигналів поруч із поняттям «сигналу» використовується поняття (чи концепція) «системи». Системи описують перетворення сигналів (аналогових та цифрових). Поняття системи зручне, коли ми хочемо зв'язати два (або більше) сигнали, що виникають при їх перетворенні або реалізувати таке перетворення. Часто зручно розглядати сигнал як вихід системи, яка може бути реальною або просто зручною моделлю. Основне призначення систем – моделювання та виконання перетворень.

Акустичні сигнали зазнають різних перетворень: розповсюдження сигналу в просторі, перетворення мікрофоном акустичного сигналу в електричний, перетворення електричного сигналу. Після перетворення електричних сигналів на цифрову форму вони обробляються вже в дискретних

системах.

Система – це певне перетворення сигналу. Система перетворює вхідний сигнал x(t) у вихідний сигнал y(t). Наприклад, залежність тиску повітря в точці іноді можна розглядати як звуковий сигнал. Залежність напруги у провіднику іноді теж може представляти звуковий сигнал. Мікрофон перетворює акустичну енергію звукового сигналу на напругу.

Модель системи – математичний спосіб опису перетворення сигналів та його характеристик. Модель системи описує зв'язки між сигналами на вході та виході (надалі – входом та виходом) і може бути описана у символьній формі наступним співвідношенням:

$$Y = F\{X\}, \qquad (3.1)$$

де X – вхід системи, Y – вихід, $F\{\}$ – перетворення, яке виконується системою.

Прикладом системи є апарат мовлення, що перетворює звук коливань голосових зв'язок у звуки мови. Перелічимо основні завдання, пов'язані із застосуванням систем. Завдання аналізу. Знаючи характеристики системи, необхідно визначити вихід по входу або вхід по виходу:

$$X, F \rightarrow Y, \qquad (3.2)$$

$$Y, F \rightarrow X. \qquad (3.3)$$

Завдання ідентифікації системи. Маючи вхід та вихід, необхідно визначити характеристики системи:

$$X, Y \rightarrow F \qquad (3.4)$$

Завдання "сліпої" ідентифікації системи (blind system identification). Знаючи вихід, необхідно визначити вхід та характеристики системи:

$$Y \rightarrow F, X. \qquad (3.5)$$

Завдання синтезу. Необхідно визначити характеристики системи, що реалізує задане перетворення між входом та виходом:

$$X, Y \rightarrow F. \qquad (3.6)$$

Моделі (поряд з аналітичними виразами) є формою представлення знань

про системи. Для того самого фізичного об'єкта можна використовувати різні моделі. Будь-яка модель має характер проекції. Модель дозволяє "представити" або "замінити" систему (рисунок 3.1).

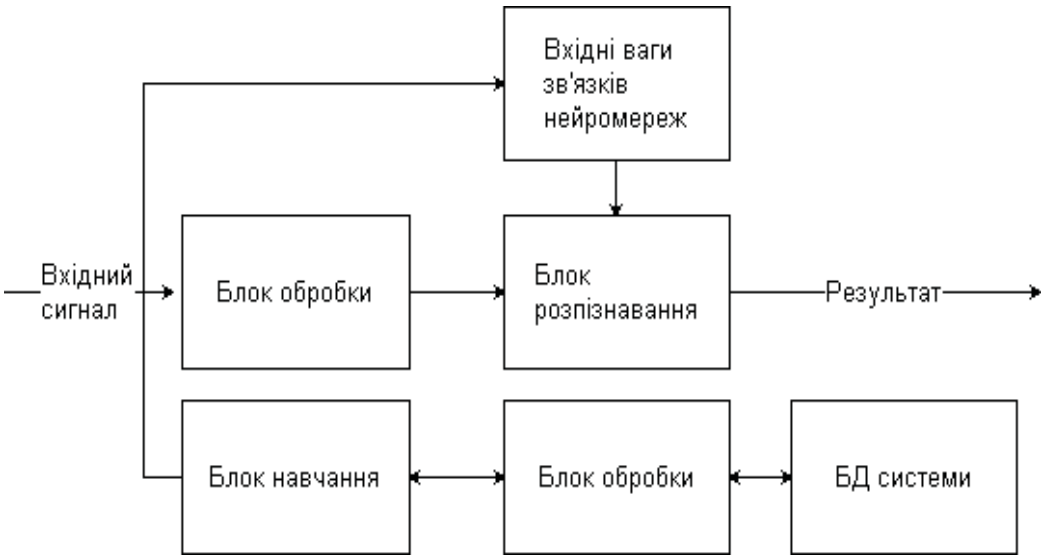


Рисунок 3.1 – Загальна схема СРМ з використанням нейромережі

Надалі, говорячи про систему, ми маємо на увазі наші уявлення про неї, тобто її модель.

Лінійна система – система, що здійснює лінійне перетворення сигналу, наприклад, електричний ланцюг з елементами, що не залежать від напруг та струмів. Лінійні системи – перше наближення (модель) перетворень сигналів. Математичне визначення лінійності:

$$y(t) = F\{a\,x1(t) + b\,x2(t)\} = a\,F\{x1(t)\} + b\,F\{x2(t)\}. \qquad (3.7)$$

Велика кількість реальних систем із перетворення сигналів можна вважати лінійними. Наприклад, мікрофон є лінійною системою (з достатнім ступенем точності), тому що якщо в нього говорити одночасно дві людини з різною гучністю, то електричний сигнал на виході буде виваженою сумою сигналів (від кожної людини окремо) на вході, а коефіцієнти будуть означати гучність розмови першої та другої людини . Однак у обробці РС широко

застосовуються і нелінійні системи. Наприклад, компресор, експандер та низка інших миттєвих перетворень є нелінійними перетвореннями. Інваріантна до тимчасового зрушення система – це система, зсув сигналу на вході якої породжує той самий відгук, лише зрушений за часом, тобто. якщо $x(t) \rightarrow y(t)$, то $x(t+\tau) \rightarrow y(t+\tau)$. Це означає, що властивості системи, наприклад мікрофонів, не змінюються у часі та не залежать від вхідного сигналу.Далі ми розглядатимемо лінійні інваріантні до зсуву системи, називаючи їх просто лінійними.

Властивості лінійних систем:

а) постійний (константа) сигнал переводиться будь-якою лінійною системою в постійний сигнал;

б) при проходженні через лінійну систему синусоїда залишається синусоїдою. Можуть змінитися лише її амплітуда та фаза (зсув у часі).

3.3 Лінійні дискретні системи

Дискретні системи – системи, що перетворюють сигнали, дискретні або за часом, або за рівнем, або за рівнем і часом (тобто цифрові сигнали). Надалі ми називатимемо дискретними системами системи, що перетворюють цифрові сигнали. Дискретні системи є дискретним аналогом безперервних систем. Співвідношення між безперервною та дискретною системами схематично представлено на малюнку .

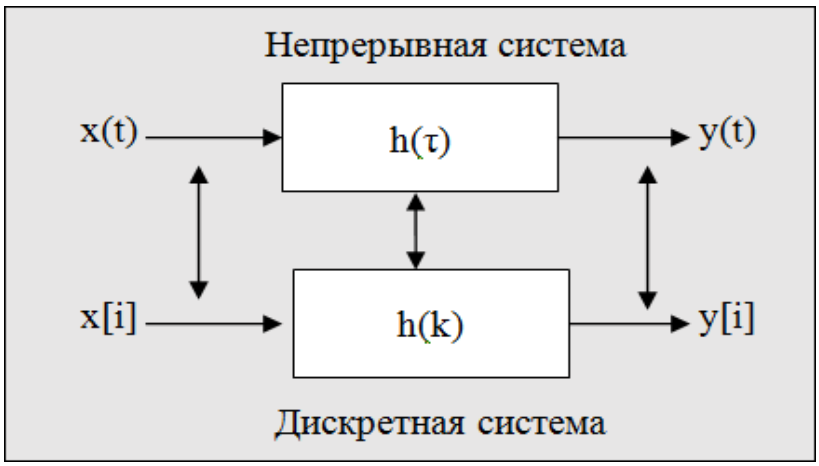


Рисунок 3.2 – Співвідношення між безперервною та дискретною системою

Під дискретизацією системи мається на увазі перетворення непре-

динамічної динамічної моделі до дискретної форми опису в різницевих рівняннях. При цьому передбачається, що моменти t = iT дискретні сигнали y[i] = y(iT) отриманої дискретної моделі з певним ступенем точності повторюють значення сигналів y(t) вихідної безперервної системи.

Дискретна модель описує перетворення дискретного вхідного сигналу x[i] у вихідний сигнал y[i]:

$$x[i] \rightarrow y[i] = F\{x[i]\}, \qquad (3.8)$$

де F{ } є дискретну систему (обчислювальний процес) перетворення вхідного сигналу x[i] у вихідний сигнал y[i].

Найпростішими дискретними системами є лінійні дискретні системи (ЛДС). Основними операціями в ЛДС є підсумовування, збільшення та затримка на один відлік (рисунок)

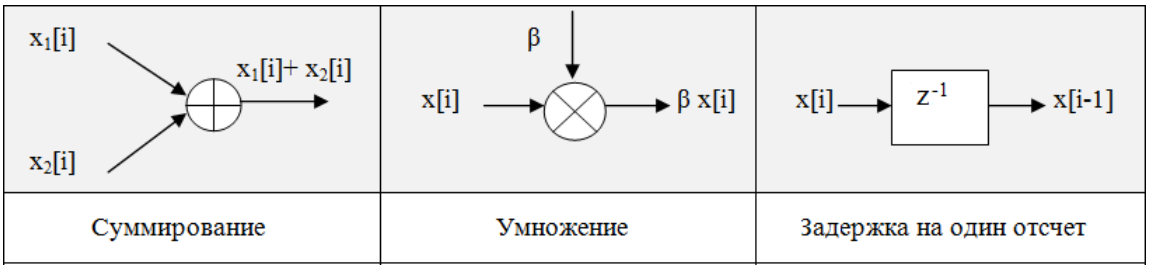


Рисунок 3.3 – Основні операції у ЛДС

z-1 – оператор затримки сигналу на один відлік

В ластивості лінійності ЛДС формулюються так само, як і для безперервних лінійних систем.

Лінійність: $y[i] = F\{ax1[i] + bx2[i]\} = a\,F\{x1[i]\} + b\,F\{x2[i]\}$. (3.9)

Інваріантність у часі: $y[i - K] = F\{x[i - K]\}$. (3.10)

3.4. Система з кінцевою імпульсною характеристикою

Систему з кінцевим числом коефіцієнтів імпульсної характеристики називають системою з кінцевою імпульсною характеристикою (КІХ). КІХ-систему (на відміну від згортки з нескінченним числом коефіцієнтів) можна представити в наступному вигляді:

$$yy[ii] = xx[ii] * h[ii] = \sum_{kk=RR1}^{RR2} xx[ii - kk]\, h[\text{i i}]. \qquad (3.13)$$

Для каузальної системи y [i] = x[i]∗h[i] = ∑k=0, + M x[i-k] h[i]. (3.14)

Для некаузальної системи y[i] = x[i]∗h[i] = ∑k= - M1 + M2 x[i-k] h[i]. (3.15)

Приклад некаузальної системи: y[i] = (x[i -1] + 2 x[i] + x[i +1]). (3.16)

Частотний опис перетворення КІХ-системи наведено нижче.

Уявімо рівняння згортки в частотній області. ДПФ згортки перетворюється на добуток ДПФ імпульсної реакції та ДПФ сигналу.

y[i]= h[i]∗ x[i] → DFT{h[i]∗x[i]} = DFT{h[i]} DFT{x[i]}. (3.17)

Розіб'ємо вхідний і вихідний сигналу на кадри кінцевого розміру (fram-ing). Застосуємо ДПФ до кадрів вхідного сигналу кінцевого розміру:

Y(f, k) = H(f) X(f, k), H(f) = Y(f, k) /X(f, k) , (3.18)

де k – індекс кадрів, f – індекс частоти.

H(f) = DFT{h[i]}, X(f, k) = DFT{x[i]}, Y(f, k) = DFT{y[i]}, (3.19)

де H(f) – передатна функція ЛДС,

H(f) = DFT{h[i]} = ∑ n=0, L-1 h[n] exp (-j2πfTn) (3.20)

Розглянемо вихідний сигнал системи:

Y(f, k) = H(k) X(k) = Re{Y} + j Im{Y} =

= (Re{H} + j Im{H}) (Re{X} + j Im{X}) =

= (Re{H} Re{X} - Im{H} Im{X}) + j (Re{H} Im{X} + Im{H} Re{X}). (3.21)

Оскільки h[n], x[n] – речові функції, то Re{H} Re{X} – симетричні, Im{H} Im{X} – антисиметричні функції, тому Re{Y} – симетрична, Im{Y} – антисиметрична. Отже вектор кадру вихідного сигналу Y[k] = IDFT{Y(f, k)}.

Послідовність кадрів вихідного сигналу перетворюється на функцію y[i]

– вихідного сигналу.

3.5. Система з нескінченною імпульсною характеристикою

У випадку перетворення сигналу в ЛДС то, можливо представлено лінійним різницевим рівнянням:

$$y[i] = \sum_{n=0, P} bn\, x[i-n] - \sum_{m=1, Q} am\, y[i-m]. \qquad (3.22)$$

Схема цього перетворення наведено малюнку 3.4.

Система має зворотний зв'язок, оскільки затриманий сигнал із виходу системи надходить назад у систему. Відгук такої системи на одиничний імпульс може тривати нескінченно. Тому такі системи називають системами з нескінченною імпульсною характеристикою (БІХ).

Розглянемо коротко опис систем з БІХ. Тимчасове уявлення: $y[i]*a[i]= b[i]*x[i]$
Частотне уявлення: $Y(f)\, A(f) = X(f)\, B(f)$

$$АЧХ: Y(f) = X(f)\, B(f)/A(f) = H(f)\, X(f),\ H(f) = B(f)/A(f) \qquad (3.23)$$

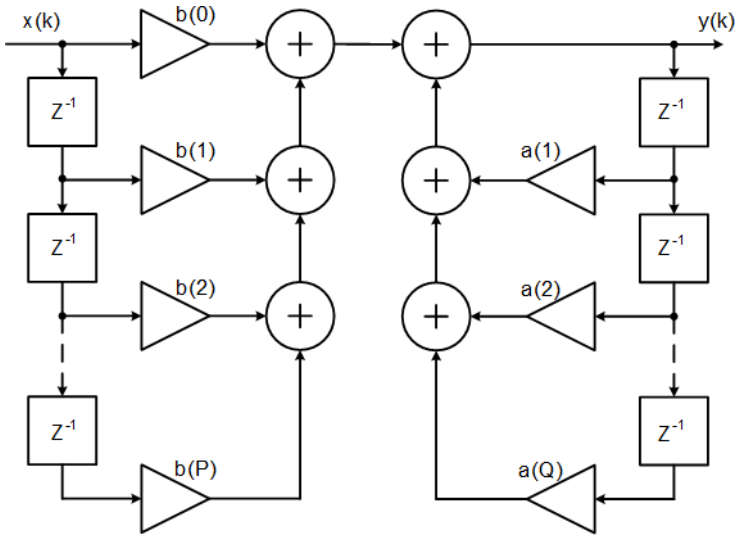


Рисунок 3.4 – Схема лінійної дискретної системи

Приклад 1

Проста БІХ-система: $y[i] = – a1\, y[i-1] + bo\, x[i] + b1\, x[i-1]$.

Приклад 2

Окремий випадок БІХ-системи (рівняння авторегресії): y[i]∗a[i] = x[i].

3.6. Спектральний опис лінійних систем з одним входом та виходом

Припустимо, що на вхід лінійної системи подається деякий стаціонарний випадковий процес x(t) з нульовим середнім. Тоді вихід y(t) матиме такі властивості:

y(t) = h(t) ∗x(t),

$$Y(f) = H(f)\, X(f) = |H(f)|\, e\, j\, \Phi(f)\, X(f), \qquad (3.24)$$

де | H (f) | – модуль передавальної функції системи (амплітудно-частотна характеристика, АЧХ), Ф(f) – фазова характеристика передавальної функції (фазово-частотна характеристика, ФЧХ).

Функції щільності спектра потужності Pxx(f), Pyy(f) та функцію крос-спектру Pxy(f) пов'язують співвідношення:

$$Pyy(f) = |H(f)|^2\, Pxx(fPxy(f) = H(f)\, Pxx(f). \qquad (3.25)$$

Якщо спектр потужності входу та крос-спектр потужності відомі, то частотна функція відгуку визначена:

$$Hxy(f) = Pxy(f)/Pxx(f). \qquad (3.26)$$

У разі ідеальної лінійної системи без шуму:

$$\Gamma^2 xy(f) = |Pxy(f)|^2 / Pxx(f)Pyy(f) =$$

$$= |\, H\, (f)\, Pxx\, (f)\, |^{\,2} / Pxx\, (f)\, |\, H\, (f)\, |^{\,2}\, Pxx\, (f) = 1. \qquad (3.27)$$

Отже, у разі лінійних систем функція когерентності досягає свого теоретичного максимуму, що дорівнює одиниці на всіх частотах. Якщо ж функція когерентності менше одиниці, то однією з можливих причин цього може бути відсутність лінійної залежності виходу від входу системи, що розглядається, тобто нелінійність системи.

3.7. Оцінка параметрів лінійних систем

Оцінку характеристик систем виконують за спостереженнями сигналу на вході та виході системи. Інші методи виконують за спостереженнями сигналу на виході системи у припущенні, що на вхід системи надходить сигнал із відомими властивостями, наприклад, білий шум. Розглянемо деякі з методів.

Оцінка передавальної функції з використанням тестового сигналу

Розглянемо лінійну систему:

$$Y(f) = |H(f)| \, e \, j\Phi(f) \, X(f). \quad (3.28)$$

Завдання полягає в оцінці амплітудної та фазової характеристики передаточної функції H(f). Оцінка за допомогою синусоїд

Метод реалізується з допомогою інструментальних засобів аналізу аналогових

сигналів. На вхід системи подається косинус частоти fo:

$$x(t) = A \cos(2\pi fo \, t). \quad (3.29)$$

Спектр вхідного сигналу має вигляд:

$$X(f) = A/2[\delta(f-fo) + \delta(f+fo)]. \quad (3.30)$$

Тоді спектр вихідного сигналу матиме вигляд:

$Y(f) = |H(f)| \, ej\Phi(f) \, A/2[\delta(f-fo) + \delta(f+fo)] =$

$|H(fo)|A/2 \, (ej\Phi(fo) + ej\Phi(-fo))$, (3.29)

а сигнал буде містити інформацію про АЧХ та ФЧХ системи:

$$y(t) = A|H(fo)| \cos(2\pi fot + \Phi(fo)). \quad (3.31)$$

Маючи в своєму розпорядженні інструмент оцінки фаз і амплітуд, можна оцінити АЧХ і ФЧХ системи [39]. Змінюючи частоту тестового сигналу, можна виміряти передачу функцію в потрібному діапазоні.

Оцінка за допомогою випадкового сигналу

Інший метод ґрунтується на цифровій обробці сигналів. На вхід системи подається широкосмуговий сигнал. Дискретні спектри вхідного та вихідного сигналів на кадрах (k) пов'язані співвідношенням:

$$Y(f, k) = H(f) \, X(f, k). \quad (3.32)$$

Крос-спектр вхідного та вихідного сигналів визначається співвідношенням:

$$Pxy(f) = \langle X^*(f, k)\, Y(f, k) \rangle = \langle X^*(f, k) H(f)\, X(f, k) \rangle =$$

$$H(f)\, Pxx(f), \qquad (3.32)$$

де () * - Символ комплексного сполучення.

Звідси може бути обчислена оцінка передавальної функції:

$$\hat{H}(f) = Pxy(f)/Pxx(f). \qquad (3.33)$$

По передавальній функції можна обчислити імпульсну характеристику системи:

$$h[i] = FT\text{-}1\{\hat{H}(f)\}. \qquad (3.34)$$

Необхідно, щоб вхідний сигнал мав спектральні компоненти у всьому діапазоні частот, що цікавить, наприклад білий шум.

Описаний метод особливо зручний для якісної експрес-оцінки АЧХ системи, якщо її вхід подавати білий шум.

Оцінка коефіцієнтів КІХ-системи по входу та виходу

Розглянемо процедуру оцінки імпульсної характеристики КІХ-системи. Запишемо відгук системи у векторній формі:

$$y[n] = hT\, X[k] = XT[k]\, h, \qquad (3.35)$$

де k – індекс кадрів, $X[k] = [x[k], x[k-1], \ldots x[k-L+1]]T$ – вектор вхідного сигналу, $h = [h[0], h[1], \ldots h[L-1]]T$ - імпульсна характеристика.

Запишемо співвідношення між вектором крос-кореляції та імпульсним відгуком:

$$\langle X[k]\, y[n] \rangle = Cxy = \langle X[k]\, XT[k]\, h \rangle = Axx\, h, \qquad (3.36)$$

де Cxy – вектор кросс-кореляції, Axx – матриця автокореляції.

Вирішуючи систему лінійних рівнянь, отримаємо оцінку коефіцієнтів імпульсного відгуку:

$$h = Axx\text{-}1\, Cxy. \qquad (3.37)$$

Оцінка коефіцієнтів авторегресійної моделі по входу та виходу

Представимо авторегресійну модель у векторній формі:

$$y[i] = AT\, Y[i] + e[i] = YT[i]\, A + e[i], \qquad (3.38)$$

де $Y[i] = [y[i-1], y[i-2], \ldots y[i-p]]T$ – вектор відліків сигналу,

A = [a [1], a [2], ... a [p]] T - Вектор коефіцієнтів.

Запишемо помилки між сигналом y[i] та виходом фільтра;

$E\{|e[i]|^2\} = < (y[i] – AT\ Y[i])^2 >.$ (3.39)

Мінімізація СКО помилки призводить до оптимального вирішення:

$<Y[i]\ y[i]> = Cy = <Y[i]\ YT[i]\ A> = Cyy\ A.$ (3.40)

Вирішуючи систему лінійних рівнянь, отримаємо оцінку БІХ-коефіцієнтів:

$$A = Cyy\ -1\ Cy. \quad (3.41)$$

Висновки до розділу

1. Описані деякі типи систем і розглядаються дві основні завдання: перше – уявлення (моделювання) аналогових систем у цифровій формі, друге – оцінка параметрів дискретних систем.

2. Розглянуті загальні питання моделювання сигналів та систем. Системи описують перетворення сигналів (аналогових та цифрових). Основне призначення систем – моделювання та виконання перетворень. Акустичні сигнали зазнають різних перетворень: розповсюдження сигналу в просторі, перетворення мікрофоном акустичного сигналу в електричний, перетворення електричного сигналу. Після перетворення електричних сигналів на цифрову форму вони обробляються вже в дискретних системах. Залежність тиску повітря в деякій точці іноді можна розглядати як звуковий сигнал. Залежність напруги у провіднику іноді теж може представляти звуковий сигнал. Мікрофон перетворює акустичну енергію звукового сигналу на напругу.

4. Автоматичне розпізнавання мовлення та розуміння природної мови

4.1 Основні принципи

Більше п’яти десятиліть дослідники мовлення працювали над створенням машини, яка зможе розпізнавати й розуміти мовленнєву мову.

Основним рушійним фактором у дослідженні стало усвідомлення того, що можливість людей спілкуватися з машинами має три потенційні основні застосування

1. Зменшення витрат: надаючи послуги, коли люди взаємодіють виключно з машинами, компанії усувають витрати на живих агентів і тим самим значно знижують вартість надання послуг. Цікаво, що в результаті цей процес часто надає людям більш природний і зручний спосіб доступу до інформації та послуг. Наприклад, авіакомпанії вже більше десяти років використовують технологію розпізнавання мовлення, що дозволяє їм інформувати своїх клієнтів про фактичний час прибуття та відправлення .

2. Нові можливості для отримання прибутку. Забезпечуючи цілодобову гарячу лінію для клієнтів і агентів з обслуговування клієнтів, компанія може дистанційно взаємодіяти зі своїми клієнтами, використовуючи природне мовлення. Прикладом такої послуги є служба AT&T «How May I Help You» (HMIHY c) яка наприкінці 2000 року автоматизувала обслуговування клієнтів для споживчих послуг AT&T.

3. Персоналізація та утримання клієнтів: надаючи низку персоналізованих послуг на основі вподобань клієнтів, компанії можуть покращити якість обслуговування клієнтів і підвищити рівень задоволеності клієнтів послугами та інформаційними системами. Дуже простим прикладом може бути персоналізація автомобільного середовища, коли автомобіль розпізнає водія за допомогою простих голосових команд, а потім автоматично налаштовує комфортні характеристики автомобіля відповідно до попередньо визначених уподобань клієнта. Хоча більша частина цієї глави буде присвячена обговоренню того, як реалізуються розпізнавання мовлення та розуміння природної мови, ми почнемо з введення концепції кола діалогу мовлення. Це діалогове коло є механізмом для всіх комунікацій між людьми. Коло діалогу мовлення зображує послідовність подій, які відбуваються між усним висловлюванням (людиною) і усою реакцією на це

висловлювання.Основна мета мови – спілкування, тобто передача повідомлень. Відповідно до теорії інформації Шеннона , повідомлення, представлене як послідовність дискретних символів, може бути кількісно визначено його інформаційним змістом у бітах та швидкістю передачі інформація вимірюється в бітах за секунду (bps). У мовному виробництві, як а також у багатьох штучних електронних комунікаційних системах, інформація, що передається, кодується у вигляді безперервно змінюваного (аналогового) сигналу, який можна передавати, записувати,маніпулюються і зрештою декодуються людиною-слухачем. Для промови фундаментальної аналогової формою повідомлення є акустична хвиля, яку називаємо мовним сигналом. Мовні сигнали можуть бути перетворені в електричні сигнали за допомогою мікрофон, що додатково керується як аналоговим, так і цифровим сигналом обробка, а потім перетворення назад в акустичну форму за допомогою гучномовця, телефонна трубка або навушники за бажанням. Ця форма обробки мови, звичайно, лежить в основі винаходу телефону Беллом. Вхідне мовлення спочатку обробляється модулем, який ми позначаємо як система ASR (автоматичного розпізнавання мовлення), єдиною метою якої є перетворення мовленнєвого введення в орфографічну послідовність слів, що є найкращою (з максимальною ймовірністю) оцінкою вимовленого висловлювання. Далі розпізнані слова аналізуються модулем розуміння розмовної мови (SLU), який намагається приписати вимовленим словам значення (у контексті завдання, яке виконує машина). Після визначення відповідного значення модуль керування діалогами визначає найбільш підходящий курс дій відповідно до поточного стану діалогового вікна та ряду встановлених кроків робочого процесу для відповідної задачі. Вжиті дії можуть бути такими ж простими, як запит на додаткову інформацію (особливо коли існує деяка плутанина або невизначеність щодо найкращої дії) або підтвердження дії, яка має бути виконана. Після прийняття рішення про дію, модуль генерації розмовної мови (SLG) вибирає текст, який потрібно промовити людині, щоб описати дію, що виконується, або запитати додаткові інструкції щодо того, як краще діяти.

Нарешті, модуль перетворення тексту в мову (TTS) промовляє висловлювання, вибране модулем SLG, і людина повинна вирішити, що сказати далі, щоб продовжити або завершити бажану транзакцію. Коло діалогового діалогу мовлення проходить один раз для кожного голосового введення та кілька разів під час типового сценарію транзакції. Вхідне мовлення спочатку обробляється модулем, який ми позначаємо як система ASR (автоматичного розпізнавання мовлення), єдиною метою якої є перетворення мовленнєвого введення в орфографічну послідовність слів, що є найкращою (з максимальною ймовірністю) оцінкою вимовленого висловлювання. Далі розпізнані слова аналізуються модулем розуміння розмовної мови (SLU), який намагається приписати вимовленим словам значення (у контексті завдання, яке виконує машина). Після визначення відповідного значення модуль керування діалогами визначає найбільш підходящий курс дій відповідно до поточного стану діалогового вікна та ряду встановлених кроків робочого процесу для відповідної задачі. Вжиті дії можуть бути такими ж простими, як запит на додаткову інформацію (особливо коли існує деяка плутанина або невизначеність щодо найкращої дії) або підтвердження дії, яка має бути виконана. Після прийняття рішення про дію, модуль генерації розмовної мови (SLG) вибирає текст, який потрібно промовити людині, щоб описати дію, що виконується, або запитати додаткові інструкції щодо того, як краще діяти. Нарешті, модуль перетворення тексту в мову (TTS) промовляє висловлювання, вибране модулем SLG, і людина повинна вирішити, що сказати далі, щоб продовжити або завершити бажану транзакцію. Коло діалогового діалогу мовлення проходить один раз для кожного голосового введення та кілька разів під час типового сценарію транзакції.

Чим точніше ASR пов’язані модулі SLU, тим більше діалогове коло виглядає як природний інтерфейс користувача для машини, і тим легше взаємодіяти та виконувати транзакції. Коло діалогу мовлення є потужною концепцією в сучасних системах розпізнавання мовлення та розмовної мови і широко використовується в ряді сучасних систем.

1. Автоматичне розпізнавання мовлення, це частини діалогового кола, де використовуються різні методи цифрової обробки сигналів. Решта частини діалогового кола є не менш важливими, але вони стосуються питань на рівні тексту, які виходять за рамки нашої праці.

4.2. Система ASR (система автоматичного розпізнавання мовлення)

Мета системи ASR полягає в тому, щоб точно й ефективно перетворити мовний сигнал у текстове повідомлення, транскрипцію вимовлених слів, незалежно від пристрою, який використовується для запису мовлення (тобто перетворювача чи мікрофона), акценту мовця чи акустики. середовище, в якому знаходиться динамік (наприклад, тихий офіс, шумне приміщення, надворі). Кінцева мета ASR, а саме розпізнавання мови так само точно і надійно, як людський слухач, ще не досягнута.

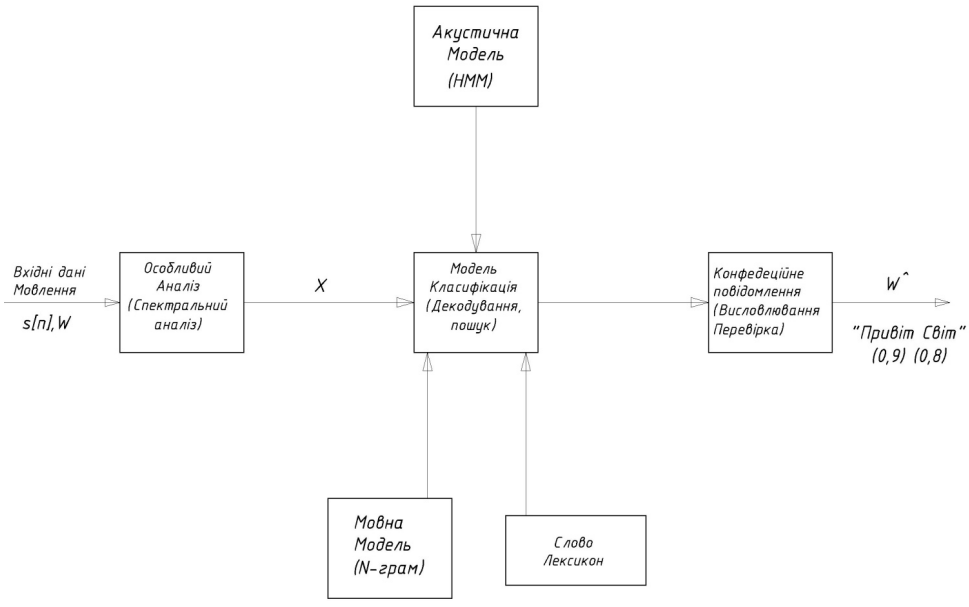
Основою більшості сучасних систем розуміння мовлення та мови є проста (концептуальна) модель генерації та розпізнавання мовлення, показана на рисунку 4.2. Передбачається, що мовець має намір висловити якусь думку як частину процесу розмови з іншою людиною або з машиною. Щоб висловити цю думку, мовець повинен скласти лінгвістично значуще речення W у вигляді послідовності слів (можливо, з паузами та іншими акустичними подіями, такими як uh’s, um’s, er’s тощо). Після вибору слів мовець посилає відповідні керуючі сигнали до артикуляційних мовленнєвих органів, які утворюють мовне висловлювання, звуки якого є тими, які потрібні для промови потрібного речення, в результаті чого формуються мовні хвилі[n].

Ми називаємо процес створення мовного сигналу з намірів мовця моделлю мовця, оскільки він відображає акцент мовця та вибір слів, які використовуються для вираження даної думки чи запиту. Етапи обробки розпізнавання мовлення показані в правій частині рисунка 4.1 і складаються з акустичного процесора, який аналізує мовний сигнал і перетворює його в набір акустичних (спектральних, тимчасових) ознак X, які ефективно

характеризують властивості звуки мови, за яким слід лінгвістичний процес декодування, який дає найкращу (максимальну ймовірність) оцінкуслова вимовленого речення, в результаті чого утворюється розпізнане речення W^.

4.3 Загальний процесс розпізнання мовлення

На рис. 4.3 показана блок-схема однієї з можливих реалізацій розпізнавача мовиВхідний мовний сигнал s[n] перетворюється на послідовність векторів ознак X X 1, X 2. . .,X T , блоком аналізу ознак (також званим спектральним аналізом). p align="justify"> Вектори ознак обчислюються покадрово з використанням методів, що обговорюються в розділах 6-9. Зокрема, для представлення короткострокових спектральних характеристик мовного сигналу широко використовується послідовність векторів крейдових крейдових



коефіцієнтів.

Рисунок 4.1 Блок-схема загальної системи розпізнавання мовленняє.

Блок-схема загальної системи розпізнавання мовлення, що складається з

обробки сигналів (аналіз для отримання набору векторів ознак, які представляють вхідне мовлення), класифікації шаблонів (щоб найкраще узгодити набір векторів ознак з оптимальним набором об'єднаних слів або звукових моделей), та оцінка впевненості (щоб отримати оцінку впевненості для кожного слова у розпізнаному вихідному рядку). щоб визначити, які слова, якщо такі є, ймовірно, будуть неправильними через помилку розпізнавання або через те, що це було поза словникового (OOV) слово (яке ніколи не можна було правильно розпізнати). Кожен з блоків на малюнку 4.1 вимагає ряду операцій з обробки сигналів і, в деяких випадках, великих цифрових обчислень. У решті цієї глави наведено огляд того, що бере участь у кожній частині обробки та обчислення на рисунок 4.1.

Нижче наведено простий приклад усного введення для розпізнаючого, що складається з фрази з двох слів і результуючого розпізнаного рядка з балами впевненості. Цей приклад ілюструє важливість і цінність оцінки впевненості.

Умовне введення: кредит, будь ласка, Розпізнаний рядок: комісія за кредит. Бали довіри: (0,9) (0,3)

На основі показників довіри (отриманих за допомогою тесту відношення правдоподібності) система розпізнавання визначить, яке слово або слова (наприклад, слово «збори» у наведеному вище прикладі), ймовірно, є помилковими (або слова OOV), і вибере відповідні кроки (у наступному діалоговому вікні), щоб визначити, чи була допущена помилка розпізнавання, і, якщо так, вирішити, як її виправити, щоб діалог міг рухатися вперед до мети завдання впорядковано та ефективно.

4.4 Базовий склад ASR

Метою системи ASR є точне і ефективне перетворення мовного сигналу в текстову транскрипцію слів, незалежно від пристрою, використовуваного для запису мови (наприклад, перетворювача або мікрофона), акценту мовлення або

акустичних характеристик. середовище, в якому знаходиться мовець (наприклад, тихий офіс, галаслива кімната, вулиця). Кінцева мета ASR, саме розпізнавання мови так само точно і надійно, як людина-слухач, ще досягнуто [30-35] .

В основі більшості сучасних систем розуміння мови та мови лежить проста (концептуальна) модель генерації та розпізнавання мови, показана на рис. 4.1. Передбачається, що той, хто говорить, має намір висловити якусь думку в процесі розмови з іншою людиною або з машиною. Щоб висловити цю думку, той, хто говорить, повинен скласти лінгвістично значущу пропозицію W у вигляді послідовності слів (можливо, з паузами та іншими акустичними явищами, такими як е-е, е-е, е-е тощо). Щойно слова обрані, промовець посилає відповідні управляючі сигнали артикуляційним органам мови, які формують мовленнєве висловлювання, звуки якого необхідні вимовлення бажаного речення, у результаті формуються мовні хвилі[n]. Ми називаємо процес створення мовної хвилі з наміру того, хто говорить моделлю мовця, оскільки він відображає акцент того, хто говорить і вибір слів, що використовуються для вираження даної думки або прохання. Етапи обробки розпізнавача мови показані у правій частині рис. 4.1 і складаються з акустичного процесора, який аналізує мовний сигнал та перетворює його на набір акустичних (спектральних, тимчасових) характеристик X, що ефективно характеризують властивості мови. звуки мови, за якими слідує процес лінгвістичного декодування, який дає найкращу (максимальну ймовірність) оцінку слів промови, внаслідок чого виходить розпізнана пропозиція.

4.4.1 Побудова системи розпізнавання мовлення.

Нижче наведено чотири кроки побудови та оцінки системи розпізнавання мовлення:

1. Виберіть завдання на розпізнавання, включаючи лексику розпізнавання слів (лексику), основні звуки мовлення для представлення

лексики (одиниці мовлення), синтаксис завдання чи мовну модель та семантику завдання (якщо є).

2. Виберіть набір функцій і пов'язану обробку сигналу для представлення властивостей мовного сигналу в часі.

3. Тренувати набір акустичних та мовних моделей.

4. Оцініть продуктивність отриманої системи розпізнавання мовлення.

Тепер ми обговоримо кожен із цих чотирьох кроків більш детально.

4.4.2 Завдання розпізнавання

Завдання розпізнавання варіюються від простих систем розпізнавання слів і фраз до розмовних інтерфейсів із великим словником до машин. Наприклад, для простого завдання розпізнавання словникового запасу з одинадцяти ізольованих цифр (/від нуля до /дев’яти/ і /oh/) як ізольованих мовленнєвих введених даних, ми можемо використовувати моделі цілих слів як основну одиницю розпізнавання (хоча звичайно можна розглядати фонемні одиниці, причому цифри складаються з послідовностей фонем). Цілі шаблони слів, якщо вони життєздатні, як правило, є найнадійнішими моделями для розпізнавання, оскільки розумно тренувати прості моделі слів практично в необмеженому діапазоні акустичних контекстів.

Використовуючи словник цифр, завданням може бути розпізнавання окремих цифр (наприклад, вибір з невеликої кількості альтернатив) або розпізнавання рядка цифр, що представляє номер телефону, ідентифікаційний код або дуже обмежений послідовність цифр, що утворюють пароль. Тому можливий ряд синтаксисів завдань навіть для такого простого словника розпізнавання.

4.4.3 Набір функцій розпізнавання

Не існує «стандартного» набору функцій, які використовувалися для

розпізнавання мовлення. Натомість різні комбінації акустичних, артикуляційних і слухових характеристик використовувалися в ряді систем розпізнавання мовлення. Найпопулярнішими акустичними характеристиками були кепстральні коефіцієнти мель-частот та їх похідні. Інші популярні вектори ознак були засновані на комбінаціях параметрів створення мови, таких як частоти формант, період тону, швидкість перетину нуля та енергія (часто у вибраних діапазонах частот). Найпопулярнішими акустично-фонетичними ознаками були ті, що описують як спосіб, так і місце артикуляції, часто оцінювані за допомогою методів нейронної обробки . Артикуляційні особливості включають функції області голосового тракту, а також положення артикуляторів у виробленні мовлення. Також були використані вектори ознак, засновані на слухових моделях. Приклади включають параметри з представлень, таких як моделі вуха Сенеффа або Ліонського вуха , або метод гістограми інтервалу ансамблю Гітца . Блок-схема обробки сигналу, що використовується в більшості сучасних систем розпізнавання мовлення з великим словником, показана на рисунку 4.3. Аналоговий мовний сигнал дискретизується і квантується зі швидкістю від 8000 вибірок/с до 20000 вибірок/сек.

Для компенсації часто використовується мережа високочастотної попередньої напруги першого порядку (1 γz -1).

для спектрального спаду мови на більш високих частотах. Попередньо підкреслений сигнал потім блокується в кадри з L вибірок, з сусідніми кадрами, віддаленими один від одного.

Типові значення дляLandR відповідають кадрам тривалістю 15–40 мс, причому найчастіше зсув кадрів становить 10 мс. Вікно Хеммінга застосовується до кожного кадру перед спектральним аналізом за допомогою методів лінійного прогнозування. Дотримуючись (за бажанням) простих методів видалення шуму, коефіцієнти предиктора, що представляють короткочасні спектр,α m , нормалізуються і перетворюються на кепстральні коефіцієнти мель-частот (mfcc)c m за допомогою стандартних методів аналізу

Перед обчисленням кепстральних похідних за часом часто використовується деякий тип видалення кепстрального зсуву. Зазвичай результуючий вектор ознак являє собою множину вирівняних кепстральних коефіцієнтів мел-кепстру c m, а також їх першого і другого порядку похідні c m[l] і 2c m[l]. Зазвичай використовують 13 коефіцієнтів mfcc, 13 першого порядку кепстральні похідні коефіцієнти та 13 похідних кепстральних коефіцієнтів другого порядку, в результаті чого вектор ознак компонента a D=39 для кожного 10-мсекундного кадру мови. Ключовою проблемою при виборі акустичних характеристик для представлення мовного сигналу є недостатня надійність багатьох відомих характеристик, пов'язаних з шумом, фоновими сигналами, передачею по мережі і спектральними спотвореннями, що призводить до невідповідності статистичних властивостей мовного сигналу. вектори акустичних ознак між етапами навчання та тестування. Блок-схема процесу вилучення ознак для вектора ознак, що складається з набору коефіцієнтів mfcc та їх першої та другої похідних систем розпізнавання.

Цей недолік надійності набору функцій часто призводить до значного погіршення продуктивності та призводить до широкого спектру запропонованих методів обробки сигналів та теорії інформації для вирішення проблем. Ряд методів, запропонованих для поліпшення цієї проблеми надійності, проілюстровано на рис. 4.3, де показано три основних методи зіставлення, а саме на рівні сигналу (за допомогою методів покращення мови), на рівні ознак (з використанням деяких типів ознак). метод нормалізації) та на рівні моделі (використовуючи будь-який метод адаптації моделі).

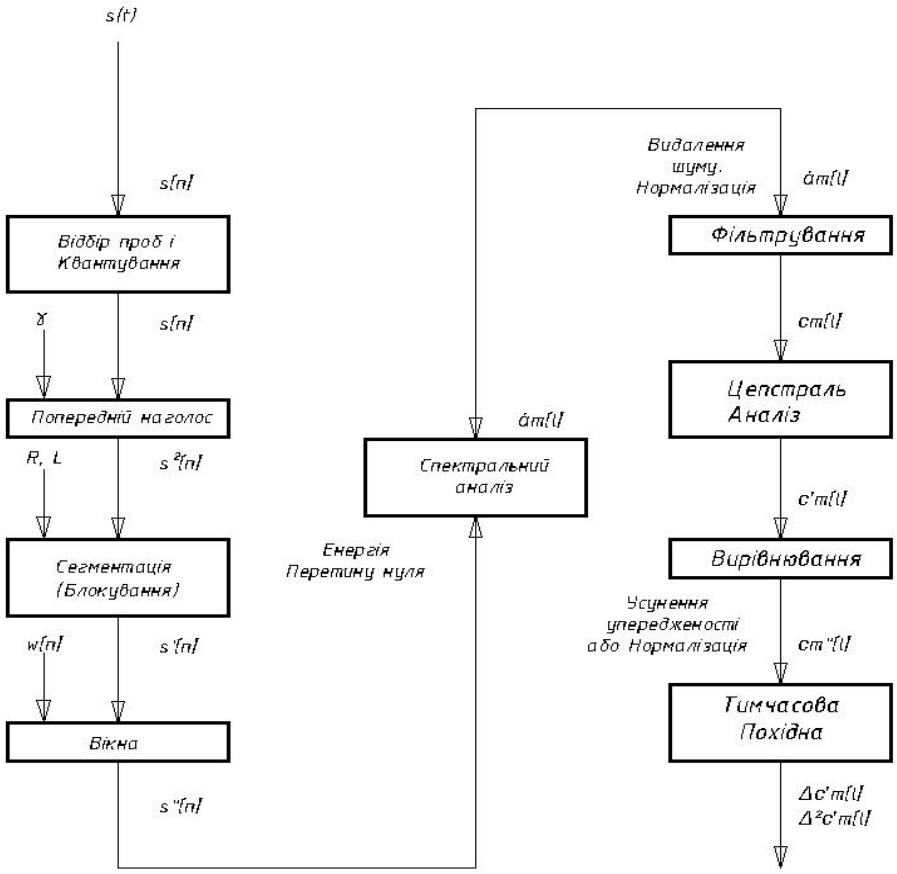


Рисунок 4.2 – Блок-схема процесу вилучення ознак для вектора ознак

4.4.4 Навчання розпізнавання

Є два аспекти моделей навчання розпізнавання мови, а саме навчання акустичної моделі та навчання мовної моделі. Навчання акустичної моделі вимагає запису кожної з одиниць моделі (повних слів, фонем) якомога більшій кількості контекстів, щоб метод статистичного навчання міг створювати точні статистичні розподіли для кожного зі станів моделі. Акустичне навчання засноване на точно маркованих мовних висловлюваннях, сегментованих відповідно до транскрипції відповідної моделі-одиниці. Навчання акустичних

моделей спочатку включає сегментацію рядків для розпізнавання модельних одиниць (за допомогою методу вирівнювання Баума-Уелча або Вітербі, як пояснюється далі в цьому розділі), а потім із використанням сегментованих висловлювань для одночасної побудови розподілів статистичних моделей для кожного стану моделей словникових одиниць. Наприклад, для простого словника з ізольованими цифрами 100 осіб, що вимовляють кожну з 11 цифр по 10 разів ізольовано, утворюють потужний навчальний набір, який забезпечує дуже точні та високонадійні моделі цілих слів для розпізнавання. Отримані статистичні моделі формують основу для операцій розпізнавання образів, що у основі системи ASR.

Для навчання мовної моделі потрібен великий навчальний набір текстових рядків, що відображають синтаксис сказаних висловлювань для завдання. Зазвичай такі навчальні набори тексту створюються машиною з урахуванням моделі граматики завдання розпізнавання. У деяких випадках у текстових навчальних наборах використовуються існуючі текстові джерела, такі як журнальні та газетні статті, стенограми телевізійних випусків новин із субтитрами тощо. В інших випадках навчальні набори для мовних моделей можуть бути створені з баз даних; наприклад, допустимі рядки телефонних номерів можуть бути створені з наявних телефонних довідників.

4.5 Тестування та оцінка продуктивності

Щоб покращити продуктивність будь-якої системи розпізнавання мовлення, має існувати надійний і статистично значущий спосіб оцінки продуктивності системи розпізнавання на основі незалежного набору тесту помічених висловлювань. Зазвичай ми використовуємо частоту помилок у словах і частоту помилок у реченнях (або завданнях) як показники продуктивності розпізнавача. Наприклад, для розпізнавання цифр із завданням розпізнавання дійсних телефонних номерів ми могли б використовувати сукупність розмовляючих з 25 нових розмовляючих, кожен із яких говорить по 10 телефонних номерів як послідовностей ізольованих (чи пов'язаних) цифр, та

був оцінити цифрову помилку. частота помилок та (обмежена) частота помилок у рядку телефонного номера.

4.6 Прихована модель Маркова

Найбільш широко використовуваним методом побудови акустичних моделей (як для фонем, так і для слів) є статистична характеристика, відома як приховані моделі Маркова (ПММ) На малюнку 4.6 показано ПММ simpleQ з 5 станами для моделювання цілого слова. Кожен стан ПММ характеризується гауссівським розподілом щільності суміші, що характеризує статистичну поведінку векторів ознак X t у станах моделі На додаток до статистичної щільності ознак всередині

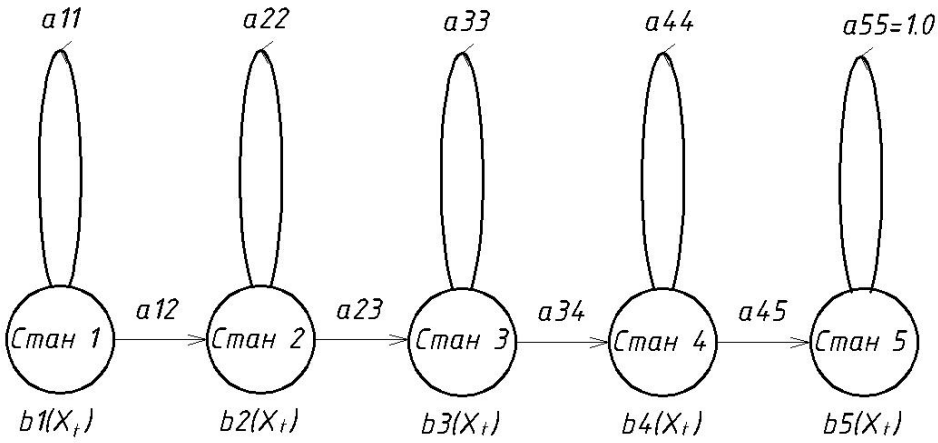


Рисунок 4.3 ПММ на основі слова з п'ятьма станами і без можливості пропуску станів, тобто aij = 0, j ≥ i + 2.

Функція щільності для стану і позначається як bi(Xt), 1 ≤ i ≤ 5 станів, НММ також характеризується явним набором переходів станів, A a ij, 1 i,j Q , для моделі aQ-стану, які визначають ймовірність здійснення переходу від стану і до statejat кожен кадр, таким чином визначаючи часову послідовність векторів ознак протягом тривалості слова. Зазвичай самопереходи стану, a ii, великі (близькі до 1,0), a переходи зі стрибком стану, a 12,a 23,a 34,a 45, в

моделі, малі (близько до 0). Модель, показана на малюнку 4.4, називається моделлю «зліва направо», оскільки стан переходить зліва направо:

$$aij = 0, j \geq i+2,\ 1 \leq i \leq Q; \quad (4.0)$$

тобто система послідовно переходить від стану до стану, починаючи з першого стану і закінчуючи останнім. Повна характеристика HMM моделі слова (абопідслова) стану aQ зазвичай записується як λ(A,B,π), з матрицею переходу станів A a ij, 1 i,j Q , густиною ймовірності спостереження стану, B b j(Xt), 1 j Q, і розподіл початкового стану,π π i, 1 i Q , при цьому π 1 встановлено на 1 (а всі інші π і встановлено на 0) для моделей «зліва направо» типу, показаного на малюнку 4.4.Щоб тренувати HMM для кожної одиниці слова (або підслова), використовується позначений навчальний набір речень (транскрибованих у слова та підслова) для керування ефективною процедурою навчання, відомою як алгоритм Баума-Велча. Цей алгоритм узгоджує кожне з різних слів (або підслова) HMM з голосовими вводами, а потім оцінює відповідні середні значення, коваріації та приріст суміші для розподілів у кожному стані моделі на основі поточного вирівнювання. Метод Баума-Велча є алгоритмом підйому на гору і повторюється до тих пір, поки не буде отримано стабільне вирівнювання моделей HMM та векторів мовних характеристик. Деталі процедури Баума-Велча виходять за рамки цієї глави, але їх можна знайти в кількох посиланнях на методи розпізнавання мови Суть процедури навчання для переоцінки параметрів моделі HMM за допомогою процедури Баума-Велча показано на рисунку 4.5. Для початку навчання використовується початкова модель HMM для кожної одиниці розпізнавання (наприклад, фонеми, слова).

Процедура навчання Баума-Велча на основі заданого навчального набору висловлювань. Величини ajk – це ймовірності переходу зі стану j у стан k, а величини {cjk, μjk, Ujk} – параметри k-го компонента суміші для моделі гаусової суміші для стану j. Початкова модель може бути обрана випадковим чином або вибрана на основі апріорного знання параметрів моделі. Ітераційний цикл — це проста процедура оновлення для обчислення ймовірностей моделі

прямого та зворотного на основі бази даних вхідної мови (навчальний набір висловлювань), а потім оптимізація параметрів моделі, щоб отримати набір оновлених HMM, при цьому кожна ітерація збільшує ймовірність журналу моделі. Цей процес повторюється до тих пір, поки не відбудеться істотного збільшення ймовірності журналу з кожною новою ітерацією Процедура навчання Баума-Велча на основі заданого навчального набору висловлювань. Величини ajk – це ймовірності переходу зі стану j у стан k, а величини {cjk, μjk, Ujk} – параметри k-го компонента суміші для моделі гаусової суміші для стану j. Це проста справа, щоб перейти зліва направо XMM для цілого слова, як показано на малюнку 4.4, до HMM для суб-одиниці слова (наприклад, фонеми).

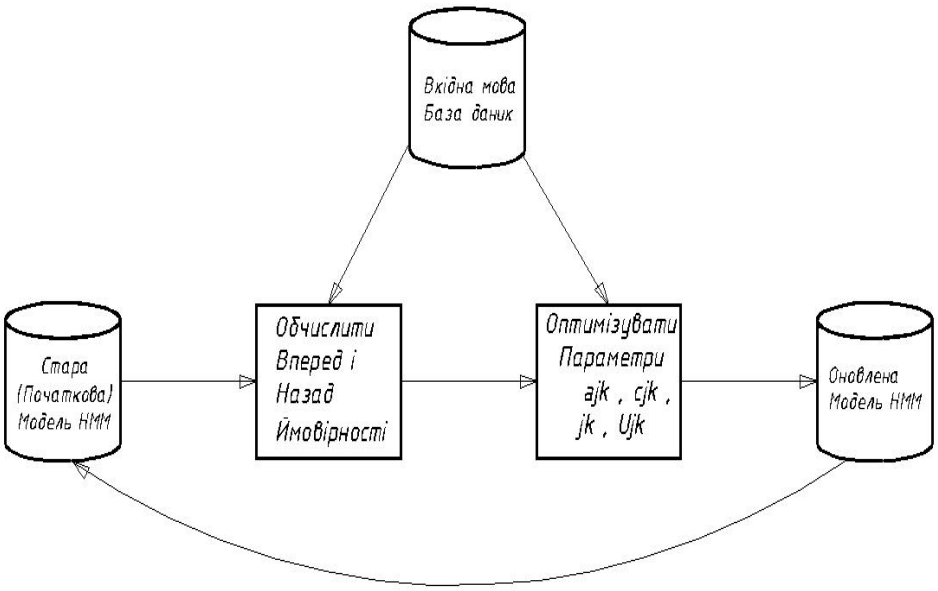


Рисунок 4.4 Процедура навчання Баума-Велча на основі заданого навчального набору висловлювань

Цей простий HMM є базовою моделлю одиниці підслова з початковим

станом, що представляє статистичні характеристики на початку звуку, в середині стан, що представляє основу звуку, і стан закінчення, що представляє

спектральні характеристики в кінці звуку. Модель слова реалізується шляхом конкатенації відповідне підслово HMM, як показано на малюнку 14.9, яке

об’єднує HMM з 3 станами для звуку /ih/ з 3-х становими HMM для звуку /z/, що дає модель слова для слова /is/ (вимовляється як /ih-z/). Загалом склад моделі слова (з моделей підсловових одиниць) уточнюється в лексиконі або словнику слів; однак, після того, як модель слова була створена, її можна використовувати майже так само, як і моделі цілого слова, для навчання та оцінки рядків слів для максимізації ймовірності в рамках процесу розпізнавання мовлення. Тепер ми готові завершити картину для вирівнювання послідовності M-моделей слів, W1, W2, . . . , WM, з послідовністю векторів ознак, X = {X1, X2, . . . , XT}. Вектори відображаються вздовж горизонтальної осі, а конкатенована послідовність станів слів — уздовж вертикальної осі. Оптимальна процедура вирівнювання визначає точну послідовність найкращої відповідності (у сенсі максимальної ймовірності) між стани моделі слова та вектори ознак, такі, що перший вектор ознак, X1, вирівнюється з першим станом у моделі першого слова, а останній вектор ознак, XT, вирівнюється з останнім станом у M-й моделі слова. (Для простоти ми показуємо кожну модель слова як Q = 5-стан HMM, але очевидно, що процедура вирівнювання працює для моделі будь-якого розміру для будь-якого слова, за умови обмеження, що загальна кількість векторів ознак T перевищує загальну кількість станів моделі, так що кожен стан має принаймні один вектор ознак, пов’язаний з цим станом.) Процедура отримання найкращого вирівнювання між векторами ознак і станами моделі ґрунтується на використанні процедури статистичного вирівнювання Баума-Велча (у якій ми оцінюємо ймовірність кожного шляху вирівнювання та додаємо їх, щоб визначити ймовірність рядка слів), або процедури вирівнювання Вітербі, для якого ми визначаємо єдиний найкращий шлях вирівнювання та використовуємо оцінку ймовірності на цьому шляху як міру ймовірності для поточного рядка слів. Корисність процедури вирівнювання заснована на простоті оцінки ймовірності будь-якого шляху вирівнювання за допомогою

процедур Баума-Велча або Вітербі.

4.7 Проблема пошуку

Ключова проблема полягає в тому, що потенційний розмір простору пошуку може бути астрономічно великим (для великого словникового запасу та високої складності мовних моделей), що вимагає надмірної обчислювальної потужность для вирішення евристичними методами. На щастя, завдяки використанню методів із польової теорії автоматів скінченних станів розвинулися методи мережі кінцевих станів (FSN), які зменшують обчислювальне навантаження на порядки величини, тим самим забезпечуючи точні рішення з максимальною імовірністю в обчислювально здійсненні часи, навіть для дуже великих проблеми розпізнавання мовлення .

Основна концепція датчика FSN проілюстрована на рисунку 4.7, де показано мережу вимови слів для слова /дані/. Кожна дуга на діаграмі стану відповідає фонемі в мережі вимови слів, а вага є оцінкою ймовірності того, що дуга використовується у вимові слова в контексті. Ми бачимо, що для слова /data/ існує чотири варіанти вимови, а саме [разом з їх (оцінюваними) ймовірностями вимови]:

1./D/ /EY/ /D/ /AX/—імовірність 0,32

2./D/ /EY/ /T/ /AX/—імовірність 0,08

3./D/ /AE/ /D/ /AX/—імовірність 0,48

4./D/ /AE/ /T/ /AX/—імовірність 0,12.

Комбінований FSN з чотирьох вимових є набагато ефективнішим, ніж використання чотирьох окремих перерахувань слова, оскільки всі дуги розподіляються між чотирма вимови, а загальне обчислення для повного FSN для слова /дані/ близьке до 1. 4 обчислення чотирьох варіантів одного слова. Ми можемо продовжити процес створення ефективних FSN для кожного слова в словнику завдання (словник мовлення або лексикон), а потім об’єднати слова FSN в речення FSN з використанням відповідної мовної моделі. Крім того, ми можемо перенести процес до рівня телефонів HMM і станів HMM, що робить процес ще ефективнішим. Зрештою, ми можемо скласти дуже велику мережу

станів моделей, моделі телефонів, моделі слів і навіть фрази в набагато меншу мережу за допомогою методу зважених перетворювачів кінцевого стану (WFST), які поєднують різні уявлення мови та мови та оптимізують отриману мережу, щоб мінімізувати кількість станів пошуку (і, еквівалентно, тим самим мінімізувати кількість повторюваних обчислень). Простий приклад такої оптимізації мережі WFST наведено на малюнку. Використовуючи методи поєднання мережі (що включає композицію мережі, детермінізацію, мінімізацію
та збільшення значення) та оптимізації мережі, WFST використовує
уніфіковану

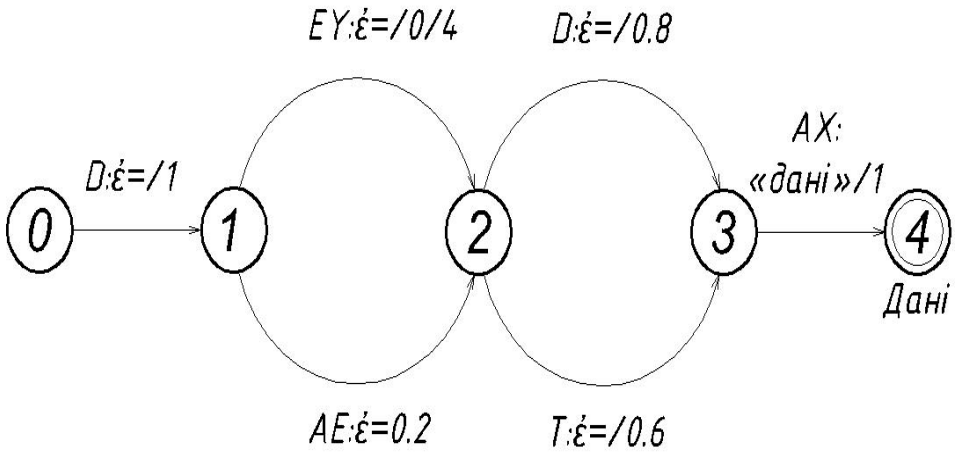


Рисунок 4.5 – Датчик вимови слів для чотирьох слів.

Математичну структуру для ефективного компіляції великої мережі в мінімальне представлення, яке легко шукати за допомогою стандартного Viterbi. Використовуючи ці методи, неоптимізовану мережу з 1022 станами

(результат перехресного добутку станів моделі, модельних телефонів, модельних слів і модельних фраз) вдалося зібрати до математично еквівалентної моделі з 108 станами, яку можна було легко шукати. оптимальний рядок слів без втрати продуктивність або точність слів.

4.8. Проста система ASR: розпізнавання ізольованих цифр

Щоб проілюструвати ідеї, представлені в цьому розділі, розглянемо реалізацію простої ізольованої системи розпізнавання цифр зі словником з десяти цифр (/від нуля/ до/дев'ять/) плюс альтернатива /oh/ для /нуль/, щоб загалом було розпізнано

11 слів. Першим кроком у реалізації розпізнавання цифр є навчання набору акустичних моделей для цифр. Для цього простого словника ми використовуємо моделі цілих слів, тому нам не потрібен лексикон слів або мовна модель, оскільки під час кожної спроби розпізнавання вимовляється лише одна цифра. Для навчання HMM цілого слова для цифр нам потрібен навчальний набір з а достатньох кількісті повторень кожної вимовної цифри, щоб мати можливість тренувати надійні та стабільні акустичні слова HMM. Як правило, п’яти маркерів кожної цифри достатньо для системи, що навчається оратором (тобто система, яка працює лише для мовлення мовця, який навчав систему). Для незалежних від динаміків розпізнавача значно більше кількість лексем кожної цифри необхідна для повної характеристики змінності наголосів, динаміків, перетворювачів, мовного середовища тощо. Зазвичай для навчання надійних моделей слів використовується від 100 до 500 лексем кожної цифри. Тренування слова HMM для кожної з цифр використовує метод Баума-Велча і дає набір слів HMM, по одному для кожної з 11 цифр, які потрібно розпізнати.

Після того, як мовний сигнал s[n] був ізольований, він перетворюється на послідовність векторів ознак mfcc, X. Потім виконується обчислення ймовірності для кожної з V = 11-значних моделей, (λ1) до (λV), що дає оцінки ймовірності цифр P(X|λ1) до P(X|λV). Кожен із показників імовірності цифр використовує декодування Вітербі для оптимального вирівнювання функції

вектор, X, із станами моделі HMM, а також для обчислення оцінки ймовірності (або еквівалентно логарифмічної ймовірності) для оптимального вирівнювання. Останнім кроком у системі розпізнавання є вибір максимальної оцінки імовірності та пов’язування індексу слова максимальної ймовірності з розпізнаною цифрою. Оцінки ймовірності завжди негативні та зменшуються.

Використовуючи HMM з 5 станами для кожної цифри, ми бачимо, що кадри з 1 по b1 − 1 тестової послідовності узгоджені зі станом моделі 1, тоді як кадри b1 до b2 − 1 тестової послідовності, узгодженої зі станом моделі 2, тощо. Зрештою, ми бачимо як кожен із п'яти станів моделі узгоджено з кадрами голосового введення, і ми це бачимоми можемо майже ідентифікувати звуки, пов’язані з кожним станом. Наприклад, ми це бачимо стан 5 узгоджено з зупинкою звуку /K/ та кінцевого /S/, тоді як стани 1 вирівнюється насамперед з початковим звуком /S/ і, нарешті, стан 4 узгоджується з голосним /IH/ та початкова частина стопового приголосного /К/.

Подібні графіки можна побудувати для тестового висловлювання /шість/ порівняно з Моделі HMM для всіх інших слів у словнику. Правильне розпізнавання відбувається, якщо оцінка імовірності журналу є найбільш негативною для еталонної моделі HMM для слова /шість/ as порівняно з тестовими векторами ознак для усного слова /шість/. Інакше –твердхення виникає помилка.

4.9. Оцінювання роботи розпізнавачів мовлення.

Ключовою проблемою в розробці системи розпізнавання мовлення (і розуміння мови) є eval визначення продуктивності системи. Для простих, ізольованих систем розпізнавання слів, наприклад розпізнавання цифр з попереднього розділу, метою продуктивності є просто частота помилок системи. Для більш складної системи, яку може бути використано для диктантів, є три типи помилок, які необхідно враховувати визначити загальну продуктивність системи. Три типи помилок, які можуть виникнути наведені

нижче:

1. Вставлення слів, коли розпізнається більше слів, ніж фактично вимовлено;

найпоширеніші вставки слів – короткі, функціональні слова, такі як «а», «ще», «ти» тощо, і найчастіше вони виникають під час пауз чи коливань у мовленні.

2. Заміни слів, де замість слова було розпізнано неправильно сказане слово; найпоширеніші помилки заміни слів є для схожих слів але іноді випадкові слова

викликають помилки заміни.

2. **Видалення слова**, коли вимовлене слово просто не розпізнається і розпізнавання не містить альтернативного слова замість нього. Видалення слів зазвичай відбуваються, коли два вимовлені слова розпізнаються як одне (зазвичай довше за тривалістю) слово,в результаті як помилка заміни слів, так і помилка видалення слова трапляється доволі часто.

3. Заснована на умові однакового зважування всіх трьох типів помилок у словах (що часто не найкраще робити, оскільки зазвичай помилки у функціональних словах значно менш шкідливі для розуміння речень, ніж помилки у змістовних словах). Загальний коефіцієнт помилок слів для більшості завдань розпізнавання мовлення, WER, формально визначається як:

$$WER = \frac{\mathrm{NI+NS+ND}}{\mathrm{W}} \tag{4.2}$$

де NI - кількість вставок слів, NS - кількість замін слів,

ND – кількість видалень слів, а W – кількість слів у реченні.

Виходячи з наведеного вище визначення частоти помилок слів, продуктивність низка систем розпізнавання мовлення та природної мови (тема ми далі в цій главі) Показники помилок слів були виміряні в ряді дослідницьких сайтів у всьому світі протягом більш ніж десятиліття, з найранішими результатами про розпізнавання рядків цифр, отримане в кінці 1980-х і на початку 1990-х років, і виміряні останні результати розмовного матеріалу природною мовою початку 2000-х років . Видно, що для

словникового запасу з 11 цифр слово помилка ставки дуже низькі (0,3% для дуже чистого середовища запису для TI (Texac Instruments) підключені цифри бази даних [204]), але коли рядки цифр промовляються у галасливому середовищі торгового центру частота помилок слів зростає до 2,0%, а коли вбудована в розмовне мовлення (система AT&T HMIHY), помилка слова показник значно зростає до 5,0%, що свідчить про недостатню надійність розпізнавання система на шум та інші фонові збурення. Коефіцієнт помилок для ряду завдань оцінки системи DARPA.

Дослідницька спільнота DARPA працювала над вищевказаним набором завдань більше ніж десятиліття (1988–2004) і прогрес, досягнутий у покращенні рівня помилок слів для кількох завдань розпізнавання мовлення.

Продуктивність людини, як правило, дуже важко виміряти, тому результати, показані на цьому малюнку, є більш спрямованими, ніж точними.

Однак очевидно, що люди перевершують машини для розпізнавання мовлення в 10-50 разів. Отже, має бути зрозуміло, що дослідження розпізнавання мовлення має пройти довгий шлях, перш ніж машини перевершать людей у завданнях розпізнавання мовлення.

Однак слід зазначити, що є багато випадків, коли ASR (і мова розуміння) система може надати кращу послугу, ніж людина. Одним із таких прикладів є розпізнавання 16-значного номера кредитної картки, що вимовляється як безперервний рядок цифр. Людині-слухачеві було б важко відстежити такий довгий (здавалося б випадковий) рядок із 16 цифр, але машина не матиме проблем із відстеженням того, що було сказано, і використання реальних обмежень кредитної картки, щоб забезпечити по суті бездоганна точність розпізнавання цифр.

Висновки до розділу

1. Окреслено основні компоненти сучасного розпізнавання мовлення і система розуміння природної мови, яка використовується в системі

голосових діалогів.

2. Показана роль обробки сигналів у створенні надійного набору функцій для розпізнавача та роль статистичних методів у наданні можливості розпізнавачу розпізнавати слова вимовленого вхідного тексту, а також значення, пов'язане з розпізнаною послідовністю слів.

3. Показано, як менеджер діалогів використовує значення накопичені з поточних, а також попередніх голосових введень для створення відповідної реакції (а також потенційно вжиття відповідних дій) на людину запит(и) і, нарешті, як завершуються частини синтезу SLG і TTS діалогу. Після вдосконалення методики, результати розрахунку стали найбільш точно збігатися з експериментальними даними, отриманими на основі випробувань зі зв'язними текстами; так по формантній заваді і рожевому шуму вони збіглися на 90%. Розбіжність по білому шуму: 20-30%.

ВИСНОВКИ:

1. Досліджено історію виникнення та розвитку систем розпізнавання мови.

2. Показано, що за сучасними тенденціями завдання є дуже актуальним.

3. Розглянуто основні ознаки, за якими можна класифікувати системи розпізнавання людської мови і те, як ці ознаки можуть впливати на роботу системи.

4. Наведений порівняльний аналіз декількох найбільш відомих та кращих існуючих систем. Окремо було досліджено синтезатори мови, а саме моделі синтезу мови та виконано порівняльний аналіз існуючих платформ.

5. Розглянуто крок за кроком архітектуру розпізнавання мови. В першому етапі виділення ознак було детально розписано алгоритм знаходження мел-частотних кепстральних коефіцієнтів (MFCC) для заданого фрейму. Наступний етап декодера включає в себе наведені три моделі:

акустична, мовна та словник, які об'єднані разом.

6. Детально оглянуто сучасні методи та алгоритми розпізнавання мовлення (Dynamic Time Wrapping, приховані марківські моделі, нейронні мережі). Особливу увагу було приділено методу, заснованого на нейронні мережах, а саме детально розглянуто процедуру навчання «класичного варіанту» багатошарової мережі.

7. Описані деякі типи систем і розглядаються дві основні завдання: перше – уявлення (моделювання) аналогових систем у цифровій формі, друге – оцінка параметрів дискретних систем. Також розглянуті загальні питання моделювання сигналів та систем. Системи описують перетворення сигналів (аналогових та цифрових). Основне призначення систем – моделювання та виконання перетворень. Акустичні сигнали зазнають різних перетворень: розповсюдження сигналу в просторі, перетворення мікрофоном акустичного сигналу в електричний, перетворення електричного сигналу. Після перетворення електричних сигналів на цифрову форму вони обробляються вже в дискретних системах. Залежність тиску повітря в деякій точці іноді можна розглядати як звуковий сигнал. Залежність напруги у провіднику іноді теж може представляти звуковий сигнал. Мікрофон перетворює акустичну енергію звукового сигналу на напругу.

8. Окреслено основні компоненти сучасного розпізнавання мовлення і система розуміння природної мови, яка використовується в системі голосових діалогів. Показана роль обробки сигналів у створенні надійного набору функцій для розпізнавача та роль статистичних методів у наданні можливості розпізнавачу розпізнавати слова вимовленого вхідного тексту, а також значення, пов'язане з розпізнаною послідовністю слів.

9. Показано, як менеджер діалогів використовує значення накопичені з поточних, а також попередніх голосових введень для створення відповідної реакції (а також потенційно вжиття відповідних дій) на людину запит(и) і, нарешті, як завершуються частини синтезу SLG і TTS діалогу. Після вдосконалення методики, результати розрахунку стали найбільш точно

збігатися з експериментальними даними, отриманими на основі випробувань зі зв'язними текстами; так по формантній заваді і рожевому шуму вони збіглися на 90%. Розбіжність по білому шуму: 20-30%.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. **Алдошина И.А.** Музыкальная акустика /И. А. **Алдошина, Р. Приттс -** С-Пб.: Композитор, 2006. - С. 717.

2. **Алдошина И. А.** Основы психоакустики / И. А. Алдошина – М.: Оборонгиз, 2000. - С. 154.

3. **Акустика.** Справочник / под ред. М.А. Сапожкова. - М.: Радио и связь, 1989.-С. 336.

4. **Кривонос Ю.Г.** Структура**,** свойства, характеристики объектов и элементов синтеза речи / Ю.Г. Кривонос, Ю.В. Крак, Н.Н. Шатковский // Компьютерная математика. – К.: Институт киберне- тики им. В.М. Глушкова НАН Украины. – 2006. – №1. – С. 61–69.

5. **Шатковський М.М.** Об'єктно-елементна модель подання текстової інформації для задачі конка- тенативного сегментивного синтезу української мови / М.М. Шатковський // Штучний інтелект. – Донецьк: ІПШІ МОН і НАН України. – 2008. – № 4. – С. 796–802.

5. **Лобанов Б.М.** Компьютерный синтез и клонирование речи / Б.М. Лобанов, Л.И. Цирульник. – Минск: Білорус. наука.– 2008. – 316 с.

6. **Шелепов В.Ю.** Структурная классификация слов русского языка. Новые алгоритмы сегментации речевого сигнала, распознавания фонем и их клас- сов / В.Ю. Шелепов, А.В. Ниценко // Искусственный интеллект. – Донецк: ИПИИ МОН и НАН Украины. – 2006. – № 4. – С. 679–690.

7. **Вінцюк Т.К**. Автоматичний озвучувач українських текстів на основі фонемно-трифонної моделі з використанням природного мовного сигналу / Т.К. Вінцюк, Т.В. Людовик, М.М. Сажок, Р. Селюх // Пр. 6-ї Всеукр. міжнар. конф. «Оброблення сигналів і зображень та розпізнавання образів» (УкрОбраз’2002). – К.: УАсОІРО. – 2002. – С. 79 – 84.

8. **Ладошко О. М**. Дослідження акустичних особливостей

вокалізованих пауз спонтанної мови / О. М. Ладошко // Міжнародна науково-технічна конференція «Штучний інтелект. Інтелектуальні системи ШІ-2011», 19-23 вересня 2011 р.: тези доп., Т. 3. – Донецьк, 2011. – С. 82-86.

9. **Ладошко О. М.** Дослідження характеристик вокалізованих пауз спонтанної мови / О. М. Ладошко // Акустичний симпозіум «Консонанс 2011», 27-29 вересня 2011 р.: тези доп. – К. – 2011. – С.188-193.

10. **Ладошко О. М.** Моделювання виділювача частоти основного тону для досліджень спонтанного мовлення / О. М. Ладошко // Міжнародна науково-технічна конференція "Моделювання і комп'ютерна графіка", 5-8 жовтня 2011 р.: тези доп. – Донецьк, 2011. – С. 304-308.

11. **Ладошко О. М**. Дослідження впливу характеристик телефонного каналу зв'язку на надійність розпізнавання фонем / О. М. Ладошко // ІІІ Міжнародна науково-технічна конференція студентів, аспірантів та молодих вчених. Інформаційні управляючі системи та комп'ютерний моніторинг (ІУС і КМ-2012), 16- 18 квітня, 2012р.: тези доп. – Донецьк. – 2012. – С. 143-148

12. Сравнительный анализ некоторых методов оценки разборчивости речи/[**Гавриленко О.В., Дидковский В.С., Продеус А.Н.**] / Акустический симпозиум «КОНСОНАНС - 2007».

13. **Гавриленко О.В., Дидковский В.С., Продеус А.Н.** Расчет и измерение разборчивости речи при малых отношениях сигнал-шум // Электроника и связь. Тематический выпуск ч 1, 2007.

14. **Григорьев С.В.** Оптимизированная по спектру шумовая помеха // Защита информации Конфидент. — № 4. — 2003.

15. **Дворянкин С.В., Макаров Ю.К., Хорев** А.А. Обоснование критериев эффективности защиты речевой информации// Защита информации. Инсайд. — С. Петербург.: 2007. — № 2 — С. 18 — 25.

16. **Железняк В.К., Макаров Ю.К., Хорев** А.А. Некоторые методические подходы к оценке эффективности защиты речевой информации// Специальная техника.

17. **Иванов В.М., Хорев** А.А. Способ и устройство формирования «речеподобных» шумовых помех // Вопросы Защиты информации. — № 4. — 1999.

18. **S. Kumar, M. Rao**, Design Of An Automatic Speaker Recognition System Using MFCC, Vector Quantization And LBG Algorithm // International Journal on Computer Science and Engineering, 2011. Vol. 3, No. 8.

19. **Покровский Н.Б.** Расчет и измерение разборчивости речи. - М.: Связьиздат, 1962.- С. 390.

20. **Продеус А.Н.** О некоторых особенностях развития объективных методов измерений разборчивости речи // Электроника и связь 2-й тематический выпучк «Электроника и нанотехнологии», 2010.

21. **Рашевский Я. П., Каргашин В. Л.** Обзор зарубежных методов определения разборчивости речи.- Специальная техника, № 3, 4, 5, 6 2002 . № 1 за >003.

22. **Сапожков М.А.** Речевой сигнал в кибернетике и связи. - М.: Связьиздат, 1963.-С. 472.

23. **Сапожков М.А.** Звукофикация помещений. Проектирование и расчет. - М.: Связь. 1979.-С. 144.

24. **Сапожков М.А., Михаилов В.Т.** Вокодерная связь. - М: Радио и связь. 1983. - С. 246.

25. **Трушин В.А.** К вопросу об оценке разборчивости речи // Проблемы информационной безопасности государства, общества, личности: - Материалы 9-й Всероссийской научно-практической конференции. Томск, 2007. С. 115-119.

26. **Vintsiuk T.K.** The analysis, pattern recognition and semantic interpretation of the speech signals. – Kiev : Naukova Dumka, 1987. – 263 p.

27. **M. Gales, S. Young.** The Application of Hidden Markov Models in Speech Recognition. // Foundations and Trends in Signal Processing. – 2007. – №1(3).

28. **T. Vintsiuk, M. Sazhok.** Multi-Level Multi-Decision Models in ASR. // Proc. of the 10th International Workshop "Speech and Computer" (SPECOM'2005). – Patras, 2005. – P. 69-76.

29. **Robeiko V.V., Sazhok** M.M. Multi-level multi-decision model transformation orthographic text to phonemic// Artificial Intelligence. – № 4'2011.

30. **Sazhok N.N.** Clustering words for construction of a linguistic model for the automatic recognition of the speech signal // Cybernetics and Computer Science: Interagency collection of scientific papers. – Issue. 170. – Kyiv, 2012.

31. **A. Lee, T. Kawahara** // Proc. European Conference on Speech Communication and Technology (EUROSPEECH). – 2001.– P. 1691-1695.

32**. Tychkov, A. Yu**. The software solutions of the problems of the biomedical information pro- cessing / A. Yu. Tychkov // Модели, системы, сети в экономике, технике, природе и об- ществе. – 2013. – № (5). – С. 114–116.

33. **M. Pinola**, Speech recognition through the decades: how we ended up with Siri, [Електродний ресурс] // PCWorld, 2011.

34**. B. Johnson**, How Siri works, [Електронний ресурс] // How Stuff Works. URL:http://electronics.howstuffworks.com/gadgets/high-tech-gadgets/siri2.htm.

35. **A. Egozi**, IAI's REX robot to face first operational trial, [Электронный ресур] // Israel Defence, 2012. URL http://www.israeldefense.com/?CategoryID=411&ArticleID=1515. 4.

ДОДАТОК А. РЕФЕРАТ АНГЛІЙСЬКОЮ МОВОЮ ЗА ТЕМОЮ ДИПЛОМНОЇ РОБОТИ

Speech recognition and speech understanding systems have made their way into mainstream applications and services and they have become widely used in society. Among the leading applications of speech recognition and natural language understanding technology include the following:

Desktop Applications: including dictation, command and control of the desktop

functionality, control of document properties (fonts, styles, bullets, etc.)

Agent Technology: including simple tasks like stock quotes, traffic reports, weather, access to communications, voice dialing, voice access to directories, voice access to messaging, calendars, appointments, etc.

Voice Portals: including systems capable of converting any web page to a voiceenabled site, where any question that can be answered on-line can be answered via a voice query using protocols like VXML, SALT, SMIL, SOAP, etc.

E-Contact Services: including call centers, customer care centers (such as How May I Help You), and help desks where calls are triaged and answered appropriately using natural language voice dialogs

Telematics: including command and control of automotive features (comfort systems, radio, windows, sunroof)

Small Devices: including cell phones for dialing by name or function, PDA control from voice commands, etc. There are two broad areas where ASR systems need major improvements before they become ubiquitous in society, namely in system and operational performance. In the system area we need large improvements in accuracy, efficiency, and robustness in order to use the technology for a wide range of tasks, on a wide range of processors, and under a wide range of operating conditions. In the operational area we need better methods of detecting when a person is speaking to a machine and isolating the

spoken input from the background. We also need to be able to handle users talking over the voice prompts (so-called barge-in conditions). More reliable and accurate utterance rejection methods are needed so that we can be sure that a word needs to be repeated when poorly recognized the first time. Finally, we need better methods of confidence scoring of words, phrases, and even sentences so as to maintain an intelligent dialog with a human. As speech recognition systems and speech understanding systems become more robust to noise, background disturbances, and transmission characteristics of communication systems, they will find their way into more small devices, thereby providing a natural and intuitive way to control the operation of these devices as well as to ubiquitously and easily access and enter information