Тема 10. Системи розпізнавання текстової інформації

1.Системи оптичного розпізнавання символів.

2. Основні принципи роботи ABBYY Fine Reader.

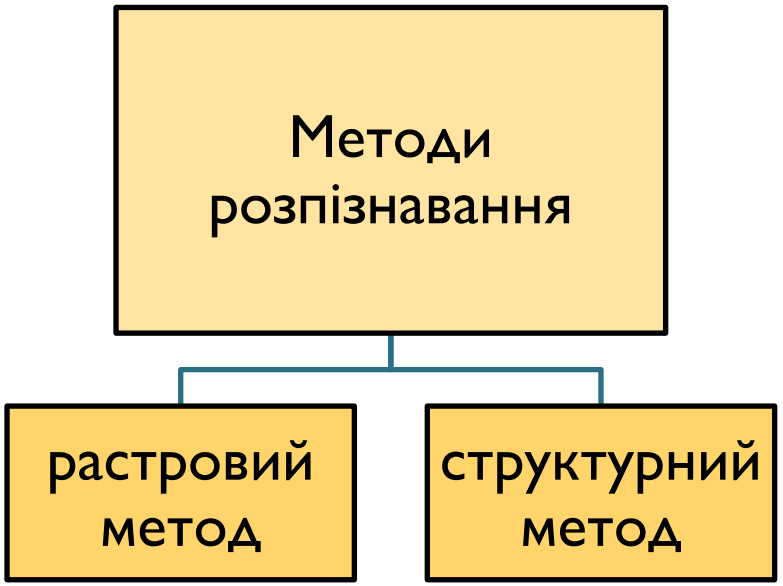
3. Багаторівневий аналіз документа (MDA).

Технологічні можливості та перспективи використання оптичних читаючих автоматів та систем розпізнавання знаку

- **Оптичне розпізнавання тексту** (англ.: *optical character recognition, OCR*) — це механічне або електронне переведення зображень рукописного, машинописного або друкованого тексту в послідовність кодів, що використовуються для представлення в текстовому редакторі. Розпізнавання широко використовується для конвертації книг і документів в електронний вигляд, для автоматизації систем обліку в бізнесі або для публікації тексту на інтернет-сторінці. Оптичне розпізнавання тексту дозволяє редагувати текст, здійснювати пошук слова або фрази, зберігати його в компактнішій формі, демонструвати або роздруковувати матеріал, не втрачаючи якості, аналізувати інформацію, а також застосовувати до тексту електронний переклад, форматування або перетворення в мовлення. Оптичне розпізнавання тексту є досліджуваною проблемою в галузях розпізнавання образів, штучного інтелекту і комп'ютерного зору.
- Системи оптичного розпізнавання тексту вимагають калібрування для роботи з конкретним шрифтом; у ранніх версіях, для програмування було необхідно зображення кожного символу, програма одночасно могла працювати тільки з одним шрифтом. Зараз найпоширеніші, так звані, «інтелектуальні» системи, що розпізнають більшість шрифтів із високим ступенем точності. Деякі системи оптичного розпізнавання тексту здатні відновлювати вихідне форматування тексту, включаючи зображення, колонки й інші нетекстові компоненти.

1. Системи оптичного розпізнавання символів

- При введенні текстової інформації у КВС, при створенні електронних бібліотек і архівів шляхом переведення книг і документів у цифровий комп'ютерний формат, при переході підприємств від паперового до електронного документообігу, при необхідності відредагувати отриманий по факсу документ використовуються *системи оптичного розпізнавання символів*. Спочатку необхідно розпізнати структуру розміщення тексту на сторінці: виділити стовпці, таблиці,



Растровий метод розпізнавання

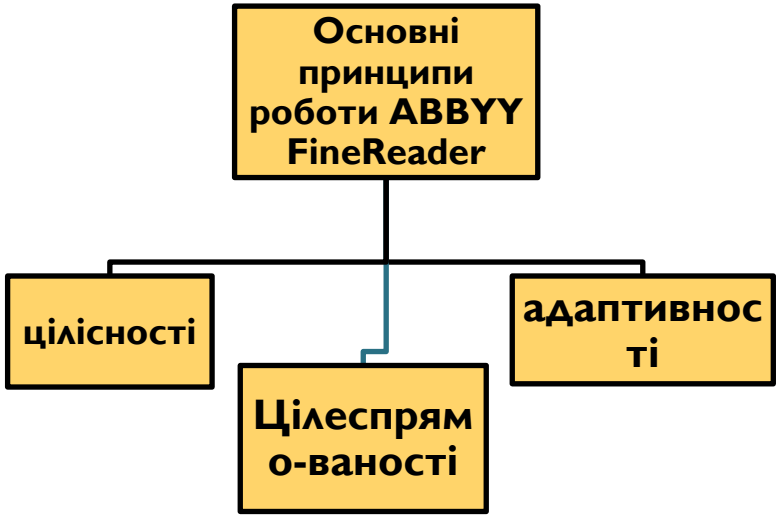
Якщо початковий документ має поліграфічну якість (достатньо великий шрифт, відсутність погано надрукованих символів або виправлень), то задача розпізнавання розв'язується методом порівняння з **растровим шаблоном.** Спочатку растрове зображення сторінки розділяється на зображення окремих символів. Потім кожний з них послідовно накладається на шаблони символів, що є в пам'яті системи, і вибирається шаблон з найменшою кількістю точок, відмінних від вхідного зображення.

Структурний метод розпізнавання

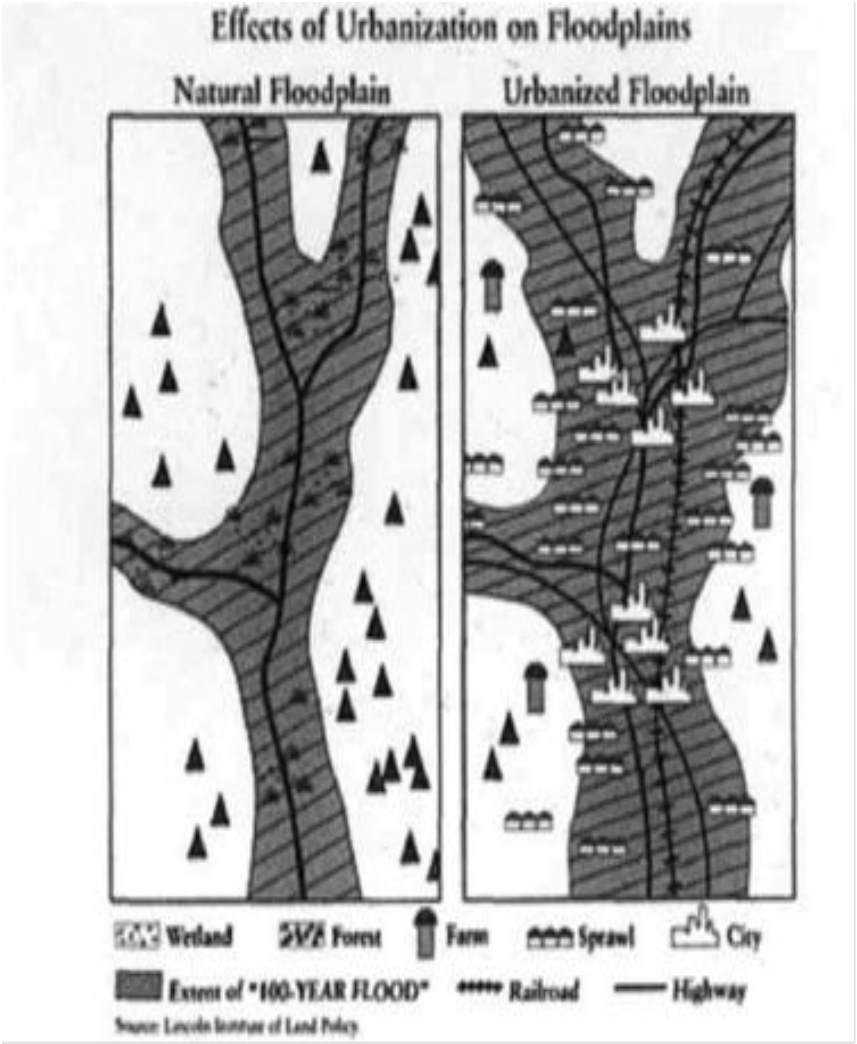
При розпізнаванні документів з низькою якістю друку (машинописний текст, факс) використовується метод розпізнавання символів за наявністю в них певних структурних елементів. Відрізків, кіцевих точок, вузлів, що визначають взаємне розташування його елементів. Наприклад, буква **«Ш»** і буква **«И»** складаються з трьох відрізків, два з яких розташовані паралельно один одному, а третій під кутом до них. Відмінність між буквами полягає у величині кутів, що має третій відрізок. Структурним методом при розпізнаванні символу зображення виділяються характерні елементи і порівнюються зі **структурними шаблонами символів**, для кожного з яких існує свій набір структурних елементів і їхнє розташування. Символ вважається розпізнаним, якщо його структура найбільше відповідає символу, що розпізнається.

2. Основні принципи роботи ABBYY FineReader

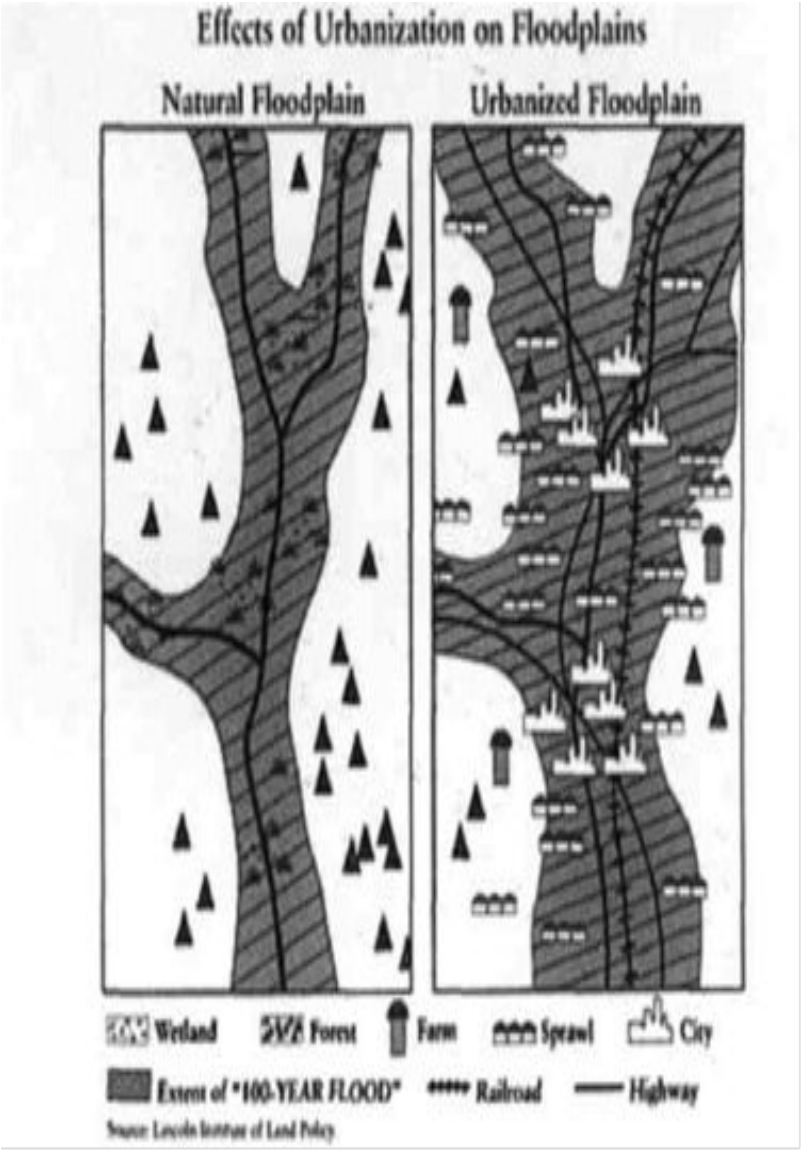
- Класична система оптичного розпізнавання працює по наступному принципу: на підставі даних про обмежений (і найчастіше фіксований) набір параметрів кожен символ порів нюється з рядом еталонів. Загалом процес виглядає так: виділивши на відсканованому зображенні об'єкти, що можуть виявитися буквами, система обчислює для кожного певний набір параметрів (таких, наприклад, як щільність чорних точок по діагоналі). Потім отримані значення по черзі порівнюються з еталонами — наборами тих же параметрів, розрахованих для відомих символів. І в залежності від того, для якого еталона різниця параметрів виявиться меншою, система прийме рішення, яким символом варто вважати виявлений об'єкт. Природно, у процесі подібного порівняння неминуче допускається деяка кількість помилок.



- **Принцип цілісності (integrity),** відповідно до якого об'єкт, що спостерігається, розглядається як ціле, що складається зі зв'язаних частин. Зв'язок частин виражається в просторових відносинах між ними, і самі частини одержують тлумачення тільки в складі передбачуваного цілого, тобто в рамках гіпотези про об'єкт.
- Приклад: ми бачимо зображення деревоподібної структури. Почато розпізнавання. Висуваються гіпотези: це або малюнок дерева, і тоді «гілки» структури відповідають гілкам, або схема автобусних маршрутів, де «гілки» позначають шляхи автобусів з різними номерами, або це карта річкової заплави, а «гілки» -



- **Принцип цілеспрямованості (purposefulness)** формулюється просто будь-яка інтерпретація даних переслідує певну мету. Відповідно до цього принципу, розпізнавання процесом висування гіпотез про цілий об'єкт і цілеспрямовану їхню перевірку.
  - Приклад (продовження): якщо зображення, яке спостерігається нами, — схема маршрутів, то на «гілках» повинні бути позначені зупинки. Якщо зображення — карта заплави, повинні бути назви рік і струмків, а також масштаб. Якщо ж це малюнок дерева, на «гілках» ймовірна наявність листів, а в основи — зображень трави або землі. Перевірка: позначень зупинок немає, листя і трави немає, у кожної «гілки» надписані назва, унизу проставлений масштаб. Підтверджено гіпотезу: це карта річкової заплави, а «гілки» відповідають руслам. Розпізнавання



- **Принцип адаптивності (adaptability)** має на увазі здатність системи до самонавчання.
- Отримана при розпізнаванні інформація упорядковується, зберігається і використовується згодом при вирішенні аналогічних задач. Перевага систем, що самонавчаються, полягає в здатності «спрямляти» шлях логічних міркувань, спираючись на раніше накопичені знання.
  - Приклад: ми бачимо нове зображення деревоподібної структури, унизу проставлений масштаб. Інформація: у минулий раз таке зображення виявилося картою, тому перш, ніж висувати інші гіпотези, варто перевірити наявність назв рік. Перевірка: назви виявлені. Розпізнавання закінчене.
- Замість повних назв цих принципів часто вживають абревіатуру IPA, складену з перших букв відповідних англійських слів. Перевати системи розпізнавання, що працює відповідно до принципів IPA, очевидні навіть неспеціалісту; саме вони здатні забезпечити максимально гнучке й осмислене поводження системи. Майже таке, як демонструють живі «розпі-навачі», створені природою.

3. Багаторівневий аналіз документа (MDA)

- сучасні OCR-програми починають розпізнавання саме з аналізу структури. Як правило, при цьому виділяють декілька ієрархічно організованих логічних рівнів. Об'єкт найвищого рівня тільки один — власне сторінка, на наступному рівні ієрархії розташовуються таблиця і текстовий блок, і так далі:

1. сторінка;
2. таблиця, блок тексту;
3. комірка таблиці;
4. абзац, картинка;
5. рядок;
6. слово, картинка усередині рядка;
7. буква (символ).

Приклад ієрархічної структури документа

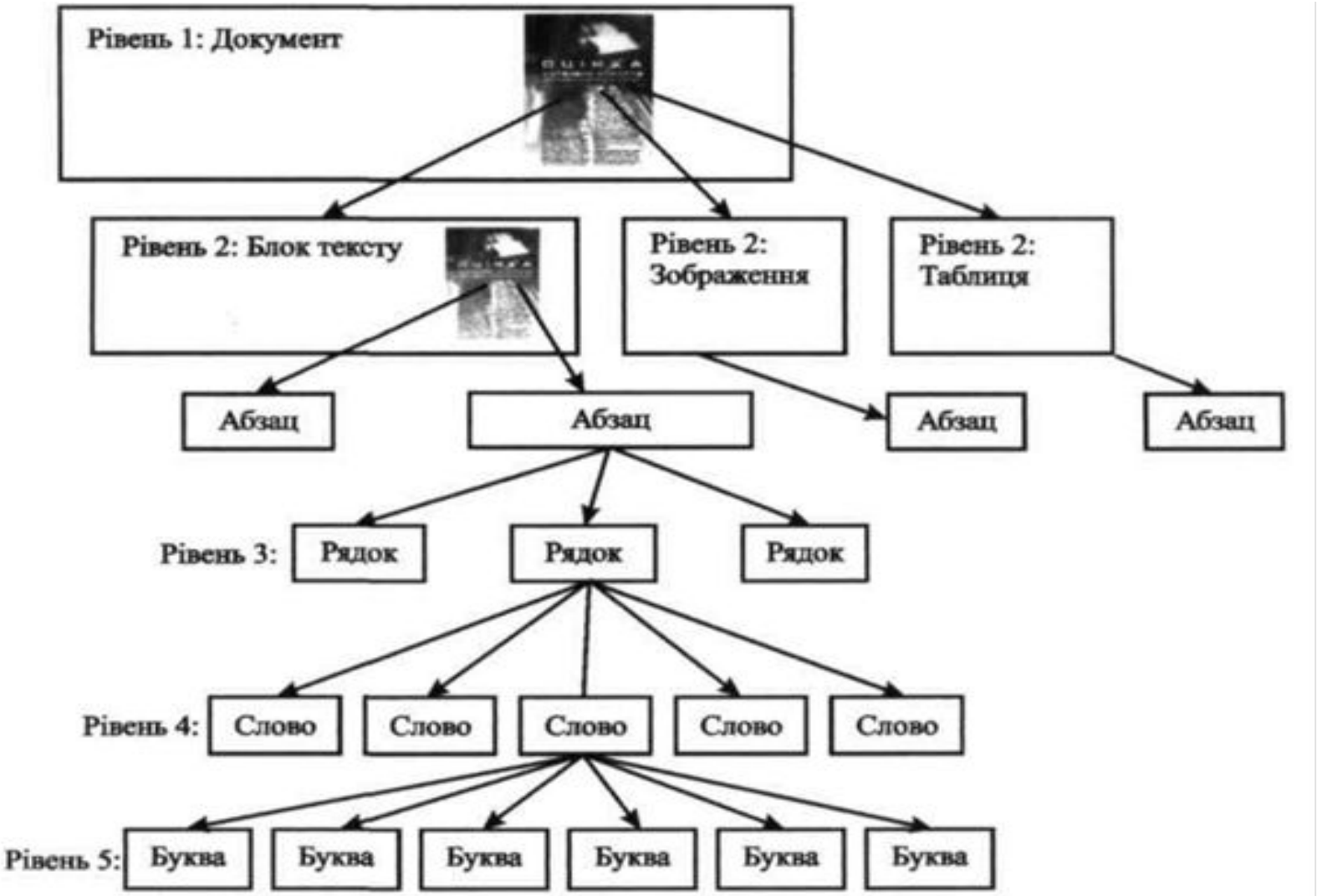


Рис. 1.3. Ієрархічна структура документу

- Зрозуміло, що будь-який високорівневий об'єкт може бути представлений як набір об'єктів більш низького рівня: букви утворять слово, слова — рядки і т.д. Тому аналіз завжди починається в напрямку зверху вниз. Програма поділяє сторінку на об'єкти, їх, у свою чергу – на об'єкти нижчих рівнів, і так далі, аж до символів. Коли символи виділені і розпізнані, починається зворотний процес — «складання» об'єктів вищих рівнів, — який завершується формуванням цілої сторінки. Така процедура називається багаторівневим аналізом документа, або MDA (multilevel document analysis).
  - Неважко бачити, що програма, що допустила помилку при розпізнаванні об'єкта високого рівня (наприклад, що переплутала абзац тексту з ілюстрацією), майже не має шансів коректно завершити процедуру — підсумковий електронний документ буде спотворений. Ризик зіткнутися з подібною ситуацією існував би і для ABBYY FineReader, якби він функціонував аналогічно більшості сучасних OCR-систем. Однак він провадить аналіз документа трохи інакше.

Алгоритм MDA

- важлива особливість використовуваного в системі ABBYY FineReader алгоритму MDA: на всіх етапах багаторівневого аналізу додана можливість зворотного зв'язку. Інакше кажучи, результати аналізу на одному з нижніх рівнів завжди можуть вплинути на дії з об'єктами більш високих рівнів. Наявність зворотного зв'язку в процедурі MDA дає можливість різко понизити ймовірність грубих помилок, зв'язаних з невірним розпізнаванням об'єктів більш високих рівнів

Процедури	***Рівень***	*Процедури*
10. Збереження документу	Документ	**1.** Виділення текстових блоків
9. Складання» документу	Блок тексту, таблиця	**2.** Виділення рядків
8. Складання» рядків	Рядок	**3.** Поділ на слова
7. Структурування гіпотез	Слово	**4.** Поділ на символи
6. Створення моделей слова	Символ	**5.** Розпізнавання символів

Рис. 1.4. Схема роботи багаторівневого аналізу документів

Висновок

- Ми коротко розглянули основні принципи роботи системи оптичного розпізнавання символів ABBYY FineReader. Як згадувалося, розпізнавання будь-якого документа провадиться поетапно, за допомогою процедури багаторівневого аналізу документа (MDA). Поділ сторінки на об'єкти нижчих рівнів, аж до окремих символів, розпізнавання цих символів і «складання» електронного документа ABBYY FineReader проводить, спираючись на принципи цілісності, цілеспрямованості й адаптивності (IPA). Такий підхід дозволяє системі забезпечити найвищу точність розпізнавання, що підтверджується результатами численних тестів, які у різний час проводилися періодичними виданнями.

Всі програми розпізнавання мови діляться на дві категорії

- програми з невеликим словниковим запасом, призначені для більшості користувачів.
  - Такі системи ідеально підходять для автоматизованого телефонного відповіді. Ці програми здатні розпізнавати декілька видів голосів, розуміти акцент і розбирати мовні зразки користувачів. Однак, управління цими програмами обмежена всього декількома зумовленими командами, наприклад, роботою з меню і управлінням з цифрами.
- програми з великим словниковим запасом, розраховані на обмежену кількість користувачів.
  - Ці системи найбільше підходять для невеликих компаній, де з програмою працюватиме тільки персонал. Але, не дивлячись на те, що ці програми працюють дуже чітко і містять кілька десятків тисяч словників, їх необхідно «підлаштовувати» під кожного користувача або під певну групу користувачів, оскільки ступінь точності може значно впасти, в разі, якщо програмою буде користуватися «не представлений» їй чоловік.
  - Системи розпізнавання мовлення, створені кілька років тому, також поділялися ще за одним критерієм – за сприйняттям мови: мова з паузами і безперервна мова. Програмі набагато легше зрозуміти окремі слова з постійною паузою між ними. Проте, більшість користувачів вважає за краще говорити зі звичайною швидкістю і не переривати свою промову постійними паузами. Тому практично всі сучасні системи здатні розуміти безперервну мову.

Сприйняття мови і її запис

- Для того щоб мова з'явилася на екрані або була сприйнята як комп'ютерна команда, комп'ютер повинен зробити декілька кроків. Коли людина говорить, вона створює коливання в повітрі. Аналого-цифровий конвертер (ADC) перетворює цю аналогову хвилю в цифрові дані, зрозумілі комп'ютеру. Під час цього процесу комп'ютер перетворює звук в цифрову форму. Потім система фільтрує переведений в цифрову форму звук і видаляє небажаний шум або перешкоди, в деяких випадках він розділяє цей цифровий звук на декілька частотних діапазонів або діапазонних частот (частота – це довжина хвилі звукових хвиль, які відчуваються людиною). Далі відбувається стандартизація звуку і регулюється його гучність. Тому системі іноді потрібен час, щоб звикнути до манери мови певного користувача. Оскільки люди постійно змінюють швидкість мови, то звук повинен бути пристосований до того, щоб швидко знаходити звукову відповідність цьому зразку з уже збережених зразків в пам'яті системи.
- ADC перетворює аналогові хвилі голосу в цифрові дані, створюючи зразки звуку. Чим вище здійснення вибірки і норми точності, тим вища якість.
- Потім сигнал ділиться на декілька сегментів, звичайно довжиною в кілька сотих частки секунди, або тисячної частки секунди, коли використовуються вибухові звуки (приголосні звуки), наприклад, англійські «p» або «t». В цьому випадку програма порівнює ці сегменти з відомими їй фонемами на зрозумілій їй мові. Фонема – це найменша одиниця мови, представлена звуками, які ми відтворюємо і з яких формується наша мова.
- Наступні дії на перший погляд здаються цілком простими, але насправді це найскладніша задача, яку намагаються вирішити більшість пристроїв розпізнавання усного мовлення. Після всіх перерахованих дій, програма починає вивчати фонеми в контексті інших фонем. Потім, як би сполучаючи фонеми в можливі слова, програма розпізнання мови порівнює їх з уже відомими словами, фразами і пропозиціями. Так програма визначає те, що говорить користувач і представляє отриману інформацію або на екрані у вигляді тексту, або сприймає її як комп'ютерну команду.

Недоліки програм розпізнавання усного мовлення

- Жодна система розпізнавання усного мовлення не може бути ідеальною. Багато з існуючих на даний момент недоліків розробники цих програм намагаються усунути в міру удосконалення технологій. Інші ж недоліки можуть усунути самі користувачі. Недоліки бувають наступними:

слабкий сигнал через навколишній шум;

- Програма повинна «чути» кожне слово чітко, і будь-який сторонній шум може порушити сприйняття голосу програмою. Шумом вважається навіть ледве чутна розмова інших працівників. Тому користувачі повинні працювати в тихій кімнаті з хорошим мікрофоном, розташованим якомога ближче до рота. До того ж звукові плати не дуже гарної якості, через які передаються сигнали з мікрофона на комп'ютер, дуже часто можуть пропускати інші електричні сигнали комп'ютерів. Тому сигнал може передаватися з гулом або шипінням.

нечіткість мови;

- Навіть сучасні системи не здатні розпізнавати одночасно мову декількох користувачів. «Тому, якщо використовувати систему розпізнавання усного мовлення під час наради або зборів, коли виступаючі часто переривають один одного, то результати можуть виявитися не зовсім очікуваними» – говорить Джон Гарофоло.

програми розпізнавання усного мовлення вимагають потужного комп'ютера;

- Статистичні моделі, що використовуються програмами для розпізнавання мови, дуже сильно завантажують комп'ютер, тому для цього процесу необхідні дуже потужні комп'ютери, здатні одночасно виконувати кілька складних функцій. По-перше, система повинна запам'ятовувати кожен свій крок, коли вона підбирає правильне слово з виголошених фонем, тому що їй може знадобитися повернутися до раніше обраного варіанту. Багато ж навіть сучасних комп'ютерів не завжди можуть опрацьовувати великі обсяги інформації, тому процесор комп'ютера може дуже повільно опрацьовувати інформацію. По-друге, самі словники цих програм займають багато місця, що ускладнює роботу процесора, якому необхідно опрацювати весь потік інформації. На щастя, все новіші технології дозволяють усунути всі ці недоліки.