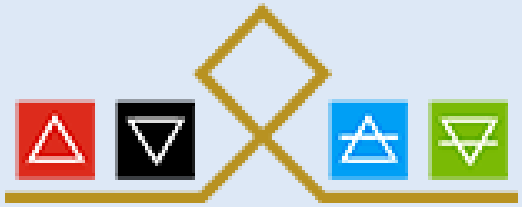
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ



Теорія розпізнавання образів та класифікації в системах штучного інтелекту

Тема №8. Розпізнавання мови

Київ - 2025

Що таке розпізнавання мови?

Система перетворення мовних сигналів у текст чи набір управляючих команд

Основне призначення систем розпізнавання мовлення

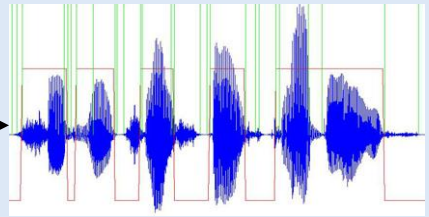
- **Управління різними пристроями за допомогою голосових команд**
- **Голосовий набір номерів**
- **Введення інформації у системи з обмеженим словником**
- **Повноцінне диктування текстів**

Основні проблеми розпізнавання

Акустична мінливість ⟹ **У різні моменти часу одні й ті ж мовні фрагменти мають різні характеристики**

Тимчасова мінливість ⟹ **У різні моменти часу одні й ті ж мовні фрагменти мають різну тривалість**

Основна схема систем розпізнавання мовлення



Моделі акустичних одиниць

Акустичний розпізнавач

о, ч', е,н',х, а,р, о,ш,ы…

База акустичних ознак

Моделі слів, фраз, речень

Лінгвістична обробка

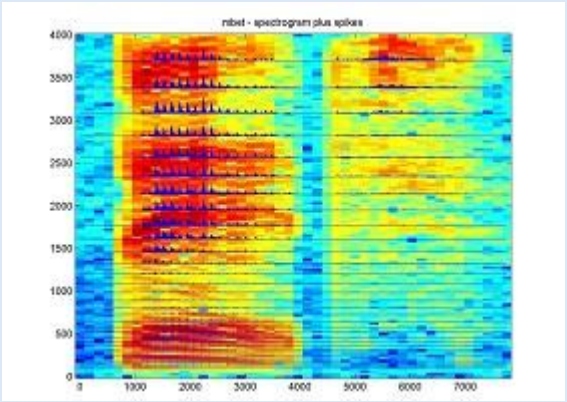
Розпізнаний орфографічний текст

“Дуже гарний сьогодні видався день……”

Правила мови

Акустичний розпізнавач

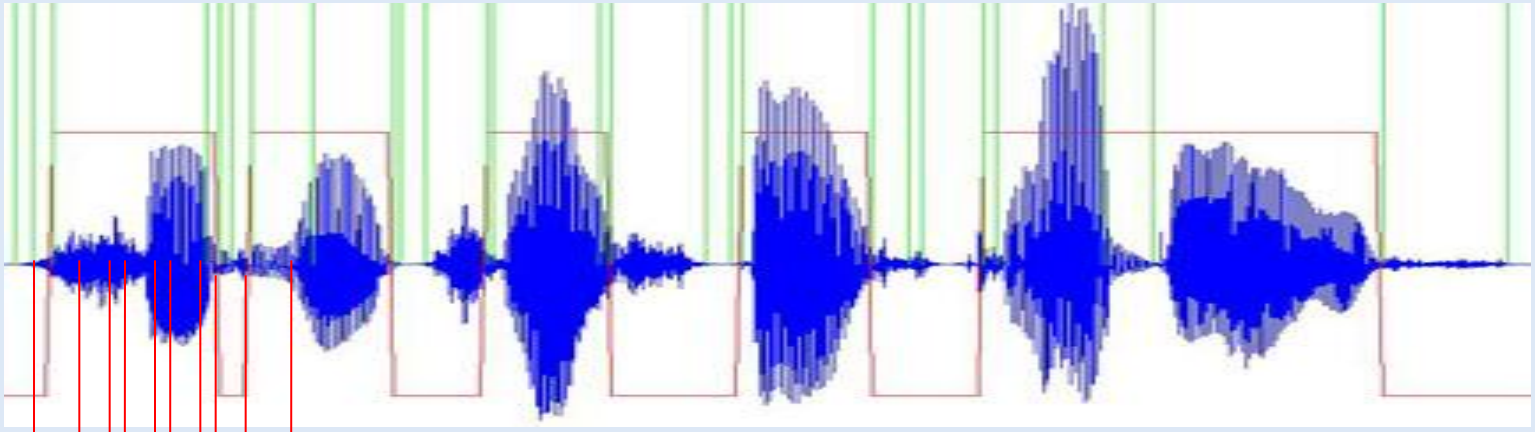
Основна мета - перетворення акустичного сигналу на послідовність акустичних одиниць, відповідних змісту вихідного сигналу



Етапи акустичної обробки

- Сегментація
- Виділення ознак
- Моделювання акустичних одиниць

Сегментація



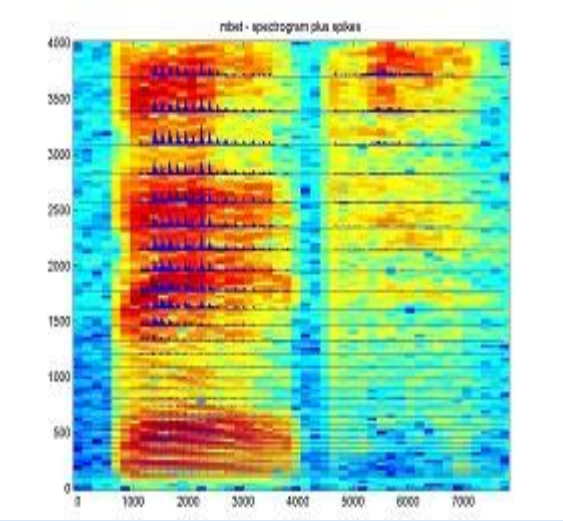
S1 S2 S3 S4 S5……..Sk

Формується послідовність перекриваються ділянок вихідного сигналу за методикою“кадр-за-кадром”

Виділення ознак

Основна мета - зіставлення кожному мовному сегменту вектора ознак

Вимоги до вектору ознак:
- інформативність
- адекватність
- стійкість
- доступність обчислення



Згладжування сигналу

↓

Спектральний аналіз

↓

Психоакустичний аналіз

↓

Кепстральний аналіз

↓

Виділення формант

↓

Дельта-параметри

↓

V1 V2 V3 V4 V5……..Vk

КОНКРЕТНИЙ ВИБІР ВЕКТОРА ОЗНАК ЗАЛЕЖИТЬ ВІД ВИРІШУВАНОЇ ЗАВДАННЯ (МОВИ, УМОВ ЗАПИСУ, т.д.)

Моделювання акустичних одиниць

Необхідно зіставити
послідовності
векторів ознак
V1 V2 V3 V4 V5……Vk
послідовність акустичних
одиниць
W1, W2, W3…….Wm

Фоне́ма — найменша неподільна звукова структурно-семантична мовна одиниця, що здатна виконувати деякі функції у мовленні. Зокрема фонема творить, розділяє і розпізнає морфеми, слова, їхні форми в мовному потоці.

Алофо́н (від грец. άλλος — «інший» + φωνή — «звук») — реалізація фонеми, її варіант, зумовлений конкретним фонетичним оточенням. Варіанти фонеми відрізняються один від одного фонетичною ознакою, а не функцією. На відміну від фонеми це не абстрактні поняття, а конкретний мовний звук.
Не зважаючи на широкий діапазон алофонів однієї фонеми, носій мови завжди спроможний їх розпізнати.
.
Дифон - лингв. сегмент мови між серединами сусідніми фонемами.

Трифон – словозаміна.

Моделі акустичних одиниць

Непараметричні моделі

- Динамічне програмування

Параметричні моделі

- Приховані марківські моделі
- Нейронні мережі
- Машина на опорних векторах
- Генетичні алгоритми
- ……

Непараметричні моделі

Зберігається копія кожної послідовності векторів ознак для кожного виразу зі словника, потім проводиться порівняння невідомого виразу з усіма збереженими копіями

Параметричні моделі

Навчається параметричну Модель для кожного виразу зі словника, потім проводиться порівняння невідомого виразу з усіма збереженими моделями

Основні проблеми розпізнавання

Дуже гарний день

Акустичний розпізнавач

о, ч', е,і,н', ух,і, р,у, о, ш, ы, д'н' …

Помилки

Заміни

Перепустки

Вставки

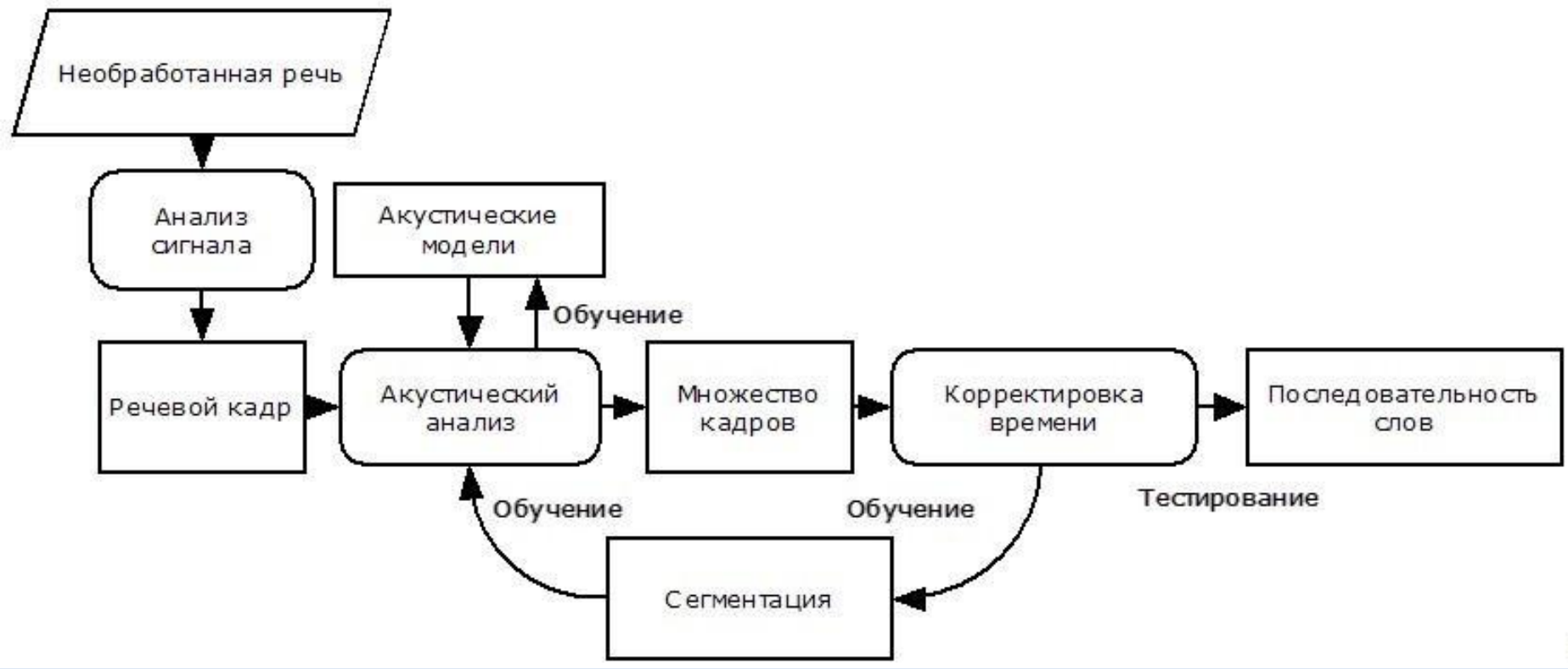
Причини помилок?

Основні причини помилок

- **Помилкова і нечітка вимова**
- **Погана дикція того, хто говорить**
- **Високий рівень сторонніх шумів**
- **Недостатнє чи погане навчання моделей**
- **Велика схожість слів словника**
- **Вимова з різною інтонацією**
- **Акцент і діалект того, хто говорить**

Системи розпізнавання мови

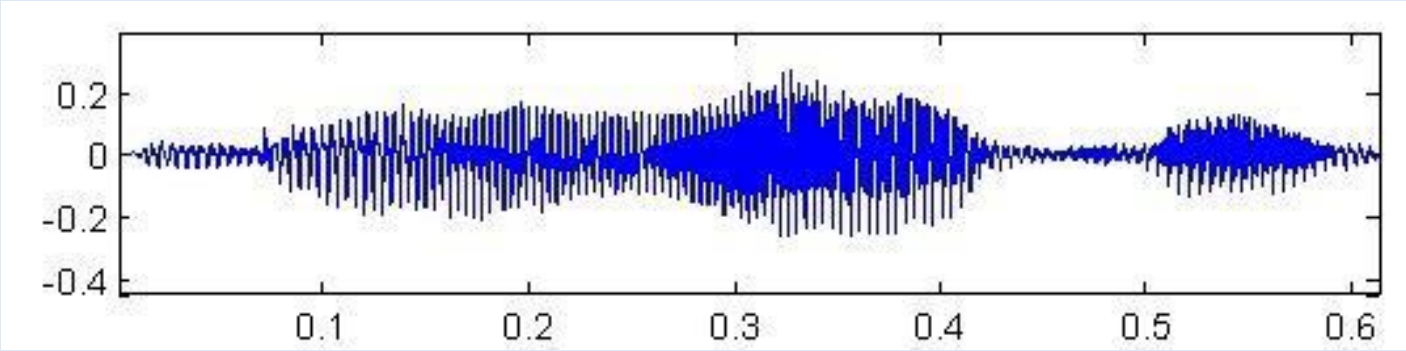
Розпізнавання мови – це багаторівневе завдання розпізнавання образів, в якому акустичні сигнали аналізуються та структуруються в ієрархію структурних елементів (наприклад, фонем), слів, фраз та речень.



Структура стандартної системи розпізнавання мовлення

Необроблена мова

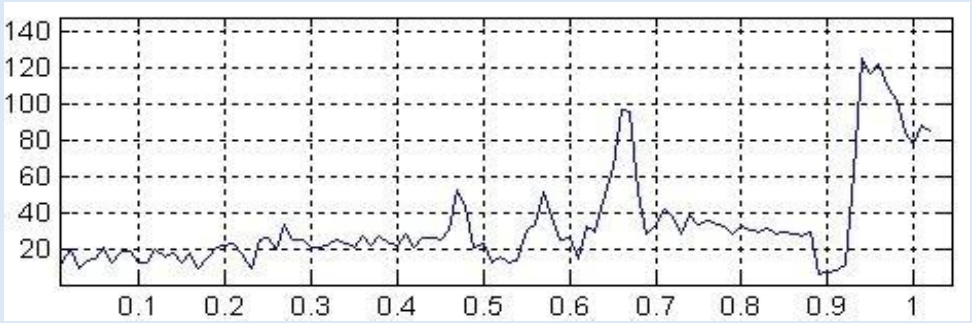
Як правило, потік звукових даних, записаний з високою дискретизацією (20 КГц при записі з мікрофона або 8 КГц при записі з телефонної лінії)



Структура стандартної системи розпізнавання мовлення

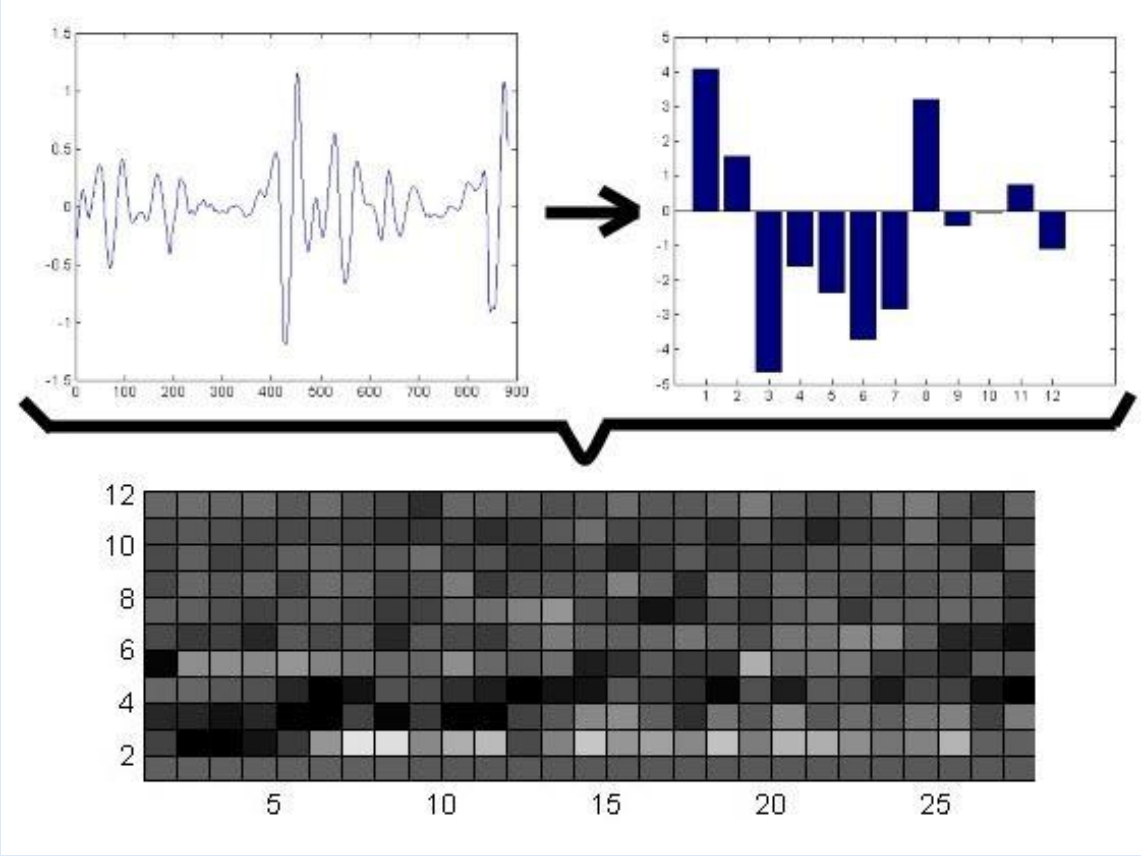
Вступниксигнал повинен бути спочатку трансформований і стиснутий для полегшення подальшої обробки. Є різні методи для отримання корисних параметрів та стиснення вихідних даних у десятки разів без втрати корисної інформації. Найбільш використовувані методи:

1. *аналіз Фур'є;*
2. *лінійне передбачення мови;*
3. *кепстральний аналіз.*



Мовні кадри

Результатоманалізу сигналу є послідовність мовних кадрів. Зазвичай, кожен мовний кадр – результат аналізу сигналу на невеликому відрізку часу (порядку 10мс.), що містить інформацію про цю ділянку (близько 20 коефіцієнтів). Для поліпшення якості розпізнавання, кадри може бути додана інформація про першу або другу похідну значень їх коефіцієнтів для опису динаміки зміни мови.



Акустичні моделі

Для аналізу складу мовних кадрів потрібен набір акустичних моделей. Найпоширеніші:

1. *шаблонна модель.*
2. *Модель станів.*

Шаблонна модель

У якості акустичної моделі виступає якимось чином збережений приклад структурної одиниці, що розпізнається (слова, команди).

Варіативність розпізнавання такою моделлю досягається шляхом збереження різних варіантів вимови одного і того ж елемента (безліч дикторів багато разів повторюють ту саму команду).

Використовується, переважно, для розпізнавання слів як єдиного цілого (командні системи).

Модель станів

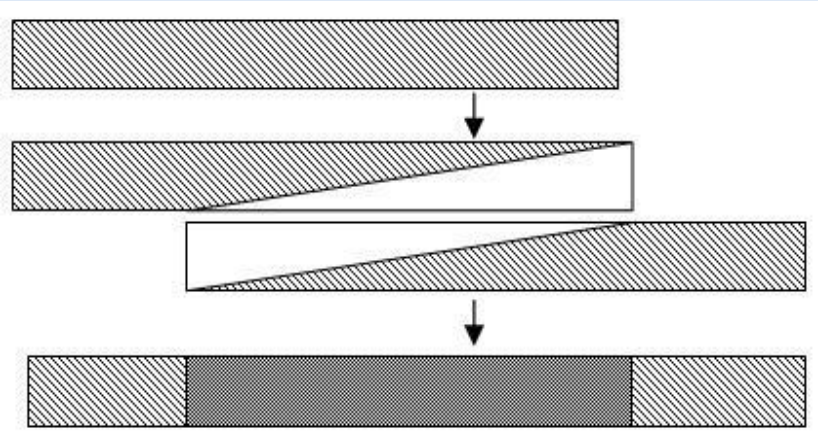
Кожне слово моделюється як послідовність станів вказують набір звуків, які можна почути у цій ділянці слова, виходячи з ймовірнісних правилах. Цей підхід використовують у більш масштабних системах.

Акустичний аналіз

Складаєтьсяу порівнянні різних акустичних моделей до кожного кадру мови і видає матрицю зіставлення послідовності кадрів та безлічі акустичних моделей. Для шаблонної моделі, ця матриця є Евклідовою відстанню між шаблонною і розпізнаваноюкадром. Длямоделей, заснованих на стані, матриця складається з ймовірностей того, що цей стан може згенерувати даний кадр.

Коригування часу

Використовується для обробки тимчасової варіативності, що виникає при вимові слів (наприклад, "розтягування" або "з'їдання" звуків).



Послідовність слів

В результаті роботи система розпізнавання мови видає послідовність (або кілька можливих послідовностей) слів, яка, найбільш ймовірно, відповідає вхідному потоку мови.

Послідовність слів

У результаті роботи система розпізнавання мови видає послідовність (або кілька можливих послідовностей) слів, яка, найбільш ймовірно, відповідає вхідному потоку мови.