Speech Recognition Model Similar to Google Speech Recognition – 15 Marks

Speech recognition, also called **Automatic Speech Recognition (ASR)**, is a technology that converts human speech into text. Systems like **Google Speech Recognition** use advanced AI, deep learning, and natural language processing to understand and process speech. A similar speech recognition model can be designed using modern AI techniques.

---

1. Introduction

- Speech recognition allows computers or machines to understand human language.
- It is used in **voice assistants** (Google Assistant, Siri, Alexa), **dictation apps**, **call centers**, and **smart devices**.
- A good speech recognition system should be:
    - Accurate
    - Able to understand different accents
    - Noise-resistant
    - Fast and real-time

Example:
When a user says, *"Open the door"*, the system converts this audio into the text *"open the door"*.

---

2. Components of a Speech Recognition Model

a) Audio Input

- Takes input from a **microphone** or **audio file**.
- Audio is converted into a **digital signal** for processing.

b) Feature Extraction

- Raw audio cannot be directly understood by AI.
- Important features are extracted using:
    - **MFCC (Mel-Frequency Cepstral Coefficients)**
    - **Spectrogram / Mel-Spectrogram**
    - **Log-Mel Filter Banks**
- These features capture **pitch, frequency, and energy** of speech.

Example:
The word *"hello"* is converted into patterns of sounds that AI can process.

c) Acoustic Model

- Converts features into **phonemes** (basic sound units).
- Uses **deep learning models** like:
    - **CNN** (Convolutional Neural Network)
    - **RNN / LSTM** (Recurrent Neural Networks)
    - **Transformer networks** (used in Whisper and Google)
- Learns:
    - How words sound
    - How tone and speed affect pronunciation
    - Differences in accents

d) Language Model

- Ensures the recognized words form **meaningful sentences**.
- Predicts the next word based on context.

Example:
Audio: *"I want to by"* → Corrected to: *"I want to buy"*

e) Decoder

- Combines outputs from the **acoustic model** and **language model**.
- Produces the final **readable text**.

---

3. Working of the Model – Step by Step

1. User speaks into a microphone.
2. Audio is captured and converted into a digital signal.
3. Features like MFCC are extracted.
4. Acoustic model identifies phonemes.
5. Language model predicts the most likely words.
6. Decoder generates the correct sentence.
7. Text is displayed to the user.

Example:
Input audio: *“Please turn on the lights”*
Output text: *“please turn on the lights”*

---

4. Example Model: Whisper (OpenAI)

- Open-source speech recognition model similar to Google’s system.
- **Features:**
    - Supports more than **90 languages**
    - Handles **long audio files**
    - Works **offline**
    - Uses **Transformer-based architecture**

    - Accurate even in **noisy environments**

How it works:

1. Converts audio into a **Mel-spectrogram**.
2. Uses a **neural network** to predict words.
3. Produces **continuous and accurate text**.

---

5. Applications of Speech Recognition

- Voice assistants (Google Assistant, Siri, Alexa)
- Automated subtitles and captions
- Dictation and voice typing software
- Call center automation
- Smart home devices
- Language translation tools
- Accessibility for disabled users

---

6. Advantages

- Hands-free operation
- Faster than typing
- High accuracy with AI
- Supports multiple languages
- Works in real-time
- Can function in noisy environments

---

7. Challenges

- Background noise may reduce accuracy
- Different accents and pronunciations
- Requires **large datasets** for training
- High computational requirements
- Privacy concerns for cloud-based systems

---

8. Conclusion

- Speech recognition systems like Google's combine **acoustic modeling, language modeling, and decoding** to convert speech into text accurately.
- Modern AI models such as **Whisper** and **Wav2Vec 2.0** show that highly accurate, real-time speech recognition can be implemented.
- These systems are widely used in **daily applications**, making speech-based interaction with machines easy and effective.