Speech to Text Model Research

1. Requirements for ASR Model

The requirements for an optimal Speech-to-Text (ASR) model are as follows:

- **Speech-to-Text Model**
    - Open Source
    - Accuracy > 80%
    - Real-time / Process Recording
    - Multilingual Support (English, Hindi, at least)

2. Open-Source ASR Models

The following open-source models are evaluated based on the above requirements:

2.1. OpenAI Whisper

- **Overview**: OpenAI Whisper is a high-accuracy, multilingual ASR model suitable for general transcription needs.
- **Advantages**:
    - High accuracy with state-of-the-art performance.
    - Multilingual support for over 50 languages.
    - Robust to noisy environments, ideal for real-world conditions.
    - Open source under MIT license, allowing free usage and customization.
    - Can handle various domains including conversational and formal speech.
    - Supports speaker diarization (separation of speakers).
    - Available in different model sizes for varying trade-offs between accuracy and performance.
- **Disadvantages**:
    - High computational cost, requiring a GPU for real-time processing.
    - Latency issues; not suitable for real-time applications with large models.
    - Limited speaker diarization capability.
    - Higher word error rate (WER) for some languages, especially low-resource ones.
    - No cloud API available, limiting ease of integration.

2.2. Coqui STT

- **Overview**: Coqui STT is an open-source, offline-capable ASR model optimized for real-time transcription.
- **Advantages**:

    - Fully open-source and free.
    - Can function offline, making it ideal for embedded systems and privacy-sensitive applications.
    - Lightweight and fast with low latency.
    - Customizable for domain-specific speech recognition.
    - Supports multiple platforms (Linux, Windows, macOS, Android).
    - Pre-trained models available for English and some other languages.
- **Disadvantages**:
    - Lower accuracy compared to Whisper and Google STT.
    - Limited multilingual support.
    - Requires model training for domain-specific accuracy.
    - Lacks built-in speaker diarization, punctuation, or capitalization features.

2.3. Vosk Speech Recognition

- **Overview**: Vosk is a lightweight, offline-capable ASR model suitable for mobile and desktop applications.
- **Advantages**:
    - Works offline, enhancing privacy.
    - Supports over 20 languages, including English, Spanish, and Hindi.
    - Cross-platform compatibility.
    - Low memory usage and pre-trained models available.
    - Speaker diarization and word alignment support.
- **Disadvantages**:
    - Lower accuracy compared to Whisper and Google STT.
    - No built-in punctuation or formatting.
    - Multilingual accuracy varies, particularly for less-resourced languages.
    - Lacks advanced AI features like sentiment analysis or context-awareness.

2.4. Facebook Wav2Vec 2.0

- **Overview**: Wav2Vec 2.0 is a self-supervised ASR model that achieves state-of-the-art accuracy with low WER.
- **Advantages**:
    - High accuracy with strong performance in low-resource languages.
    - Works offline and supports both cloud and on-premise deployment.

    - Open-source and customizable for domain-specific needs.
    - Better noise robustness than traditional ASR models.
- **Disadvantages**:
    - Requires high computational resources (powerful GPUs).
    - Slower inference speed, not ideal for real-time transcription.
    - No built-in punctuation and capitalization.
    - Requires fine-tuning for best results and domain-specific tasks.

2.5. Kaldi

- **Overview**: Kaldi is a customizable toolkit for speech recognition research, widely used in academic settings.
- **Advantages**:
    - Highly customizable with low WER when properly trained.
    - Open-source, with support for multilingual speech recognition.
    - Effective for large-scale speech data and speaker diarization.
    - Offline support and deployment flexibility.
- **Disadvantages**:
    - Steep learning curve; requires strong ASR and machine learning knowledge.
    - No pre-trained models; requires extensive training and setup.
    - High computational requirements for training large models.
    - Slower inference speed compared to lightweight models like Vosk.

2.6. SpeechBrain

- **Overview**: SpeechBrain is a modular and extensible toolkit for speech processing, including ASR and speaker recognition.
- **Advantages**:
    - State-of-the-art accuracy using deep learning models like RNNs and Transformers.
    - Supports multilingual models and self-supervised learning.
    - Open-source and free, with pre-trained models available.
    - Customizable for specific domains and applications.
    - GPU-optimized for fast training and inference.
- **Disadvantages**:
    - Slower inference speed, not ideal for real-time applications.
    - Requires machine learning expertise to set up.

    - Lacks built-in punctuation and formatting.
    - Computationally intensive with a smaller community compared to Kaldi.

3. Model for Required Data Extraction

To complement the ASR models, the following tools are considered for data extraction, particularly for Named Entity Recognition (NER) and Personally Identifiable Information (PII) extraction:

3.1. spaCy + Named Entity Recognition (NER)

- **Advantages**:
    - Free and open-source.
    - Can extract various entities like names, addresses, and organizations.
    - Suitable for general NER tasks.
- **Disadvantages**:
    - Cannot natively extract phone numbers or emails.
    - Accuracy varies for complex entity types.

3.2. Flair

- **Advantages**:
    - Developed by Facebook Research.
    - Higher accuracy than spaCy for certain entity types.
- **Disadvantages**:
    - Limited out-of-the-box functionalities.

3.3. Presidio by Microsoft

- **Advantages**:
    - Specializes in PII extraction.
    - Can detect phone numbers, emails, SSNs, and more.
- **Disadvantages**:
    - Limited support for non-PII entity extraction.

3.4. Regex-Based Approach

- **Advantages**:
    - Lightweight and fast for specific extractions like phone numbers and emails.
- **Disadvantages**:
    - Not suitable for complex NER tasks.

4. ASR Model Comparison

Feature	OpenAI Whisper	Coqui STT	Vosk	Wav2Vec 2.0	Kaldi	SpeechBrain
License	MIT (Open-source)	MPL (Open-source)	Apache 2.0 (Open-source)	Facebook AI (Open-source)	Apache 2.0 (Open-source)	Apache 2.0 (Open-source)
Pre-Trained Models	Yes	Yes	Yes	Yes	No	Yes
Languages Supported	50+	Limited	20+	Multilingual (fine-tuning needed)	Any (if trained)	10+
Real-time Processing	No (high latency)	Yes	Yes	No (slow)	Partially	No (slow)
Offline Support	Yes	Yes	Yes	Yes	Yes	Yes
Customization	Limited	Yes	Yes	Yes	Yes	Yes
Accuracy (WER - English)	2.7%	~7-10%	~10%	4.3%	5-7%	4.5-6%
Accuracy (WER - Multilingual)	3-6%	Limited	~15%	8-12%	Varies	8-12%
Inference Speed	Slow	Fast	Fast	Slow	Slow	Slow
Best For	General ASR, Multilingual	Real-time, Custom ASR	Embedded, Offline, Low-power	Self-Supervised ASR	Research, Custom ASR	Research, Advanced ASR
Hardware Requirements	GPU recommended	CPU or GPU	CPU-friendly	GPU recommended	High CPU/GPU	GPU recommended
Streaming Support	No	Yes	Yes	No	Partial	No
Speaker Diarization	No	Yes	Yes	Yes	Yes	Yes
Punctuation & Capitalization	Yes	No	No	No	No	No

Ease of Use	Easy (Pre-trained API)	Easy (Simple API)	Easy	Difficult	Complex	Medium