# Research on Various Speech Recognition Models

Speech recognition technology has advanced significantly, leveraging deep learning and AI to improve accuracy and support multiple languages. This document explores various speech recognition models in detail, discussing their functionality, advantages, limitations, features, accuracy, and reference research papers.

## Whisper by OpenAI

Accuracy: 90-95%  
Features: Multilingual (100+ languages), high accuracy, noise-robust, supports speech-to-text and translation.  
  
Whisper is an advanced ASR model developed by OpenAI. It is trained on large-scale multilingual data, making it highly robust against different accents, background noise, and low-quality audio. The model operates using transformer-based deep learning, similar to GPT architectures.  
  
🔗 Reference: https://github.com/openai/whisper

## DeepSpeech by Mozilla

Accuracy: 85-90%  
Features: Open-source, offline support, customizable training.  
  
DeepSpeech is an open-source ASR engine based on deep neural networks (RNN). It provides offline speech recognition, allowing it to function without internet access, making it suitable for privacy-focused applications.  
  
🔗 Reference: https://github.com/mozilla/DeepSpeech

## Wav2Vec 2.0 by Facebook AI

Accuracy: 92-97%  
Features: Self-supervised learning, requires less labeled data, supports low-resource languages.

Wav2Vec 2.0 is a state-of-the-art ASR model developed by Facebook AI (Meta). It learns directly from raw audio data, reducing dependence on labeled datasets. It performs exceptionally well in multilingual environments.  
  
🔗 Reference: https://huggingface.co/facebook/wav2vec2-base-960h

## Vosk

Accuracy: 80-90%  
Features: Lightweight, real-time processing, offline support, supports 20+ languages.  
  
Vosk is an efficient speech recognition toolkit that works offline. It is designed for use on mobile and embedded systems, making it ideal for real-time ASR applications.  
  
🔗 Reference: https://github.com/alphacep/vosk-api

## Kaldi

Accuracy: 85-95%  
Features: Research-grade toolkit, highly customizable, supports DNNs and HMMs.  
  
Kaldi is widely used in academia and industry for building ASR systems. It provides advanced customization options for training deep neural networks, making it suitable for large-scale applications.  
  
🔗 Reference: https://kaldi-asr.org/

## Julius

Accuracy: 75-85%  
Features: Fast, lightweight, real-time ASR.  
  
Julius is an open-source ASR engine optimized for speed. It is used in embedded systems and applications where processing efficiency is a priority.  
  
🔗 Reference: https://github.com/julius-speech/julius

## CMU Sphinx

Accuracy: 70-80%  
Features: Offline support, multiple ASR engines (PocketSphinx, Sphinx-4).  
  
CMU Sphinx is a flexible speech recognition toolkit supporting offline ASR. While its accuracy is lower than deep-learning-based models, it remains widely used in research.  
  
🔗 Reference: https://cmusphinx.github.io/

## Google Speech-to-Text

Accuracy: 95-98%  
Features: Cloud-based ASR, supports 125+ languages, real-time transcription.  
  
Google Speech-to-Text is a cloud ASR service powered by deep learning. It is one of the most accurate ASR systems available, commonly used in customer support automation and voice-controlled applications.  
  
🔗 Reference: https://cloud.google.com/speech-to-text/

## IBM Watson Speech to Text

Accuracy: 93-97%  
Features: Customizable vocabulary, supports multiple domains, cloud-based.  
  
IBM Watson Speech-to-Text is designed for enterprise applications requiring specialized terminology. It offers features such as domain adaptation and speaker diarization.  
  
🔗 Reference: https://www.ibm.com/cloud/watson-speech-to-text

# Comparison of Speech Recognition Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy (%) | Features | Offline Support | Multilingual Support |
| Whisper (OpenAI) | 90-95 | High accuracy, multilingual, noise robust | No | Yes (100+ languages) |
| DeepSpeech (Mozilla) | 85-90 | Open-source, offline support, trainable | Yes | Limited |
| Wav2Vec 2.0 (Facebook AI) | 92-97 | Self-supervised learning, low resource support | Yes | Yes |
| Vosk | 80-90 | Lightweight, real-time, offline support | Yes | Yes (20+ languages) |
| Kaldi | 85-95 | Highly customizable, research-grade toolkit | Yes | Yes |
| Julius | 75-85 | Fast, lightweight, real-time processing | Yes | Limited |
| CMU Sphinx | 70-80 | Offline support, customizable models | Yes | Limited |
| Google Speech-to-Text | 95-98 | Cloud-based, real-time, supports 125+ languages | No | Yes |
| IBM Watson Speech to Text | 93-97 | Industry-focused, customizable vocabulary | No | Yes |

# Analysis and Recommendations

Based on the comparison, the best model for a form filling using Speech Recognition :

1.Whisper (OpenAI): Best for multilingual, general-purpose ASR.

2. Vosk: Ideal for offline, real-time speech recognition.

3. Wav2Vec 2.0: Great for low-resource languages and customization.

# Conclusion

Speech recognition technology has evolved significantly, enabling high-accuracy multilingual transcription. The choice of model depends on factors such as offline capability, customization, and accuracy. For a multilingual form-filling application, Whisper and Google Speech-to-Text are the best choices, while Vosk is ideal for offline applications.