ChatGPT

Automatic Speech Recognition (ASR) Models (Newest → Oldest)

- **Massively Multilingual Speech (MMS)** – *Meta (2023):* Supports ~1,107 languages (fine-tuned on Bible readings in 1,100+ languages) [1]. Fine-tunable (open code). GitHub: Facebook Fairseq MMS. Associated with Meta's 2023 release (no formal paper yet; see Meta AI blog). Achieved about **half the WER of Whisper** on a 54-language benchmark [2] (i.e. significantly lower error) while extending ASR to 10× more languages. *Model size:* ~1 billion parameters (Transformer encoder). *Inference:* Slower than Whisper (uses ~20GB RAM; ~10× slower than optimized Whisper) [2] due to large model size and many languages. *Release:* May 2023.

- **OpenAI Whisper** – *OpenAI (2022):* A multilingual ASR supporting **98 languages** (including Bengali) [3]. Open-source (MIT License) and fine-tunable (via Hugging Face Transformers). GitHub: `openai/whisper` [3]. Described in *"Robust Speech Recognition via Large-Scale Weak Supervision"* (Radford et al., 2022). Notable for near *human-level English ASR*: Whisper-Large (~1.55B params) attains ~**9.9% WER** on diverse English test sets (approaching human transcriber ~8.8% WER) [4] [5]. For low-resource languages (e.g. Bangla), Whisper requires fine-tuning and was outperformed by specialized models in one study [6] [7]. *Model sizes:* Tiny (39M) → Large-v2 (1.55B). *Inference:* Real-time on GPU; Medium/Large models ~0.5–1× real-time on a V100. Optimized "FasterWhisper" versions achieve faster inference (at some accuracy cost). *Release:* Sept 2022.

- **AI4Bharat IndicWav2Vec** – *AI4Bharat (2022):* A **multilingual Wav2Vec 2.0** model for 40 Indic languages (incl. Bengali) [8]. Open-source and fine-tunable (Fairseq/Hugging Face) [9]. Associated paper by Chaudhary et al. (2022). Pre-trained on 17k hours across 40 languages [8]; fine-tuned ASR models released for 9 languages. Achieved state-of-art on public benchmarks (MUCS, OpenSLR etc.). For Bengali, fine-tuned IndicWav2Vec yields **16.6% WER** (13.6% with language model) on test data [10] – a strong result given prior baselines ~74% WER [11]. *Model size:* ~317M parameters (24-layer Transformer) similar to Wav2Vec2-Large. *Release:* June 2022.

- **WavLM** – *Microsoft (2021):* A **self-supervised Transformer** (24-layer) pre-trained on 94k hours for "full-stack" speech tasks (ASR, diarization, etc.). Open-source (MIT) and fine-tunable. Hugging Face: `microsoft/wavlm-base` etc. Paper: *"WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing"* (Chen et al., 2022). WavLM-Large (approx. 300M params) matches Wav2Vec2 and HuBERT on ASR, achieving ~**1.8% WER (clean) / 3.2% (other)** on LibriSpeech (comparable to Wav2Vec 2.0's 1.8/3.3% [12]). Uniquely, WavLM improved **speaker diarization** and speech separation: e.g. a 12.6% DER reduction vs. EEND on CALLHOME (from ~20% to ~17.5% DER) [13] [14]. *Inference:* real-time on GPU. *Release:* Oct 2021.

- **XLS-R (128-lingual Wav2Vec2)** – *Meta (2021):* A massively multilingual **Wav2Vec2** model pre-trained on 128 languages (436k hours) [15] [16]. Open-source (Fairseq) and fine-tunable. Paper: Babu et al., 2021 (arXiv:2111.09296). Released in **300M, 1B, and 2B-parameter** variants [15] [16]. Showed across-the-board improvements: e.g. 14–34% relative WER reduction on BABEL, CommonVoice, MLS compared to prior XLSR-53 [17] [18]. Supports many low-resource languages (likely including Bengali). *Release:* Dec 2021.

- **Wav2Vec 2.0 & XLSR-53** – *Facebook (Meta) (2020):* **Wav2Vec 2.0** is a breakthrough self-supervised ASR model (Transformer encoder with CNN feature extractor). Open-source (Fairseq) and **fine-tunable** for ASR via a CTC head [19]. Paper: *Baevski et al., NeurIPS 2020*. When fine-tuned on 960h LibriSpeech, Wav2Vec2-Large achieved **1.8% WER on test-clean, 3.3% on test-other** (with a 4-gram LM) – **state-of-the-art** at the time [19]. The cross-lingual variant **XLSR-53** (Conneau et al. 2020) was trained on 56k hours in 53 languages (including Bengali) [20] [21] with a 24-layer, 300M-param Transformer. XLSR-53 dramatically improved low-resource ASR, e.g. a 72% relative phoneme error rate reduction on CommonVoice vs prior methods [22] [23]. GitHub: Facebook Fairseq. *Release:* Oct 2020 (Wav2Vec 2.0), Dec 2020 (XLSR-53).

- **QuartzNet** – *NVIDIA (2020):* A **streamlined convolutional ASR** model (Jasper architecture variant) using 1D time-channel separable CNNs. Open-source via NVIDIA NeMo. Paper: Kriman et al. (ICASSP 2020). Notable for its small size (~**18M params**) and strong accuracy. QuartzNet-15x5 (with 15 residual blocks) achieved **~3.2% WER (test-clean) / 7.5% (test-other)** on LibriSpeech with decoding LM [24] [25] – approaching Jasper's accuracy with 10× fewer parameters. Fine-tunable on low-resource data (demonstrated via transfer learning) [26] [25]. *Release:* Oct 2019 (arXiv 1910.10261).

- **Jasper** – *NVIDIA (2019):* A deep end-to-end **CNN acoustic model** ("Just Another Speech Recognizer"). Open-source (NeMo toolkit). Paper: Li et al. (Interspeech 2019). Jasper has a 54-layer CNN with residual connections (5 blocks of 3 sub-blocks each) [27], totaling ~**332M parameters**. It achieved **<3% WER on LibriSpeech** test sets (w/ beam decoding) – a state-of-the-art result among end-to-end models without external data [28] [29]. Jasper's release underscored end-to-end models rivaling traditional hybrid systems on ASR. *Release:* Sept 2019.

- **Mozilla DeepSpeech** – *Mozilla (2017-2020):* An open implementation of Baidu's DeepSpeech (RNN-CTC model). Versions 0.6–0.9 released as open-source (TensorFlow). Fine-tunable (Mozilla provided pre-trained English model). Achieved ~**7–8% WER on LibriSpeech test-clean** with a trigram LM (around v0.7) [30]. Real-time inference was feasible on CPU – even on a Raspberry Pi 4 [31] – reflecting its light architecture. However, accuracy lagged newer Transformer models. *Model:* 5-layer bidirectional LSTM (~47M params). *Release:* 2017 (v0.1) – Feb 2020 (v0.9). Mozilla's project is now continued by Coqui STT (2021).

Speaker Diarization Models (Newest → Oldest)

- **NVIDIA Sortformer** – *NVIDIA (2023):* An *end-to-end diarization* Transformer model integrating diarization with ASR output sorting (addresses speaker permutation). Open-source (NeMo toolkit; HF model `nvidia/diar_sortformer`). Paper: Park et al. (arXiv 2024) [32]. The model uses an 18-layer Transformer (115M params) and can handle up to 4 speakers [33]. It achieved **DER 8.5%** on 3-speaker CALLHOME and **~14.8% DER on DIHARD3** (eval set) without separate clustering [34] [35] – outperforming prior diarization pipelines. Fine-tunable (requires substantial GPU). *Release:* late 2023.

- **NeMo MSDD** – *NVIDIA (2022):* A *hybrid diarization pipeline* in NeMo employing **Multi-Scale Diarization Decoder (MSDD)**. Uses a pre-trained speaker embedding model plus a transformer-based re-segmentation to handle overlap. Open-source in NeMo toolkit. (Ref: Shafey et al. 2021). Achieved strong results on phone call benchmarks (e.g. ~**8.2% DER on CallHome-2spk** condition) [36]. Fine-tunable components (speaker embedder or MSDD) with NeMo. Often used with oracle VAD. *Release:* 2021–22 (NeMo 1.7).

- **Pyannote (Speaker-Diarization Pipeline)** – *Inria (2019–2023):* A popular open-source **toolkit** for diarization (Bredin et al.). Provides pre-trained **pipelines** on Hugging Face for speaker diarization that are *fine-tunable on custom data* [37]. Current version 3.x ("Community") achieves **~10–17% DER** on various benchmarks (e.g. 17.0% on AMI, 26.7% on CALLHOME for the open model) [38] [39] – state-of-the-art among open solutions [40]. The *Pyannote* pipeline includes voice activity detection, speaker segmentation, embedding (speaker encoder), and clustering (or neural labeling). *Model sizes:* e.g. Speaker embedding model ~2.2M params (ResNet), segmentation model ~20M. *Inference:* very fast – e.g. ~**31 seconds to process 1 hour** audio on GPU (community model) [41] (~0.0086× real-time). *Release:* First version in 2020 (ICASSP) with continual updates (v3 in 2023).

- **EEND (End-to-End Neural Diarization)** – *Hitachi & NTT (2019–2021):* A **single-model diarization** approach (Fujita et al., Interspeech 2019) that directly outputs frame-level speaker activities using a BLSTM/Transformer encoder [42] [43]. Open-source (MIT); GitHub: *hitachi-speech/EEND*. Fine-tunable on diarization data. Original EEND (2-speaker BLSTM) cut DER nearly in half vs. clustering on mixes with heavy overlap. E.g. ~**20–25% DER** on CALLHOME (2-spk) vs ~40% for clustering [42]. Later *EEND-EDA* (Encoder-Decoder Attractor, Fujita+ 2020) handles variable speakers with attractor vectors, further reducing error (e.g. **15.5% DER on 2-spk DIHARD3** eval) [44]. Variants like AES *and Conformer EEND reached ~12% DER on CALLHOME in research settings* [45]. Model size: *~7M (BLSTM) up to ~20M (Transformer-EDA).* Release:* Oct 2019 (repo), improved versions in 2020–21.

- **UIS-RNN** – *Google (2019):* Stands for *Unbounded Interleaved-State RNN*, an early learned **clustering model** for diarization (Zhang et al., 2019). Open-source on GitHub. It uses precomputed speaker embeddings (d-vectors) and a small RNN to assign speakers sequentially. Fine-tunable (requires labeled dialogs). In practice, UIS-RNN combined with Google's d-vector embeddings achieved ~**12–13% DER** on NIST CALLHOME (2-speaker telephone) – comparable to x-vector clustering. It was a pioneering trainable alternative to heuristic clustering. *Model:* small RNN (≈0.5M params). Now largely superseded by end-to-end methods. *Release:* 2019.

- **Kaldi X-vector + Clustering** – *Kaldi toolkit (2018):* A classic *modular diarization* approach. Uses a pre-trained **x-vector speaker embedding** model (Snyder et al., 2018) and PLDA or cosine clustering. Open-source (Kaldi recipes) – can be adapted ("fine-tuned") by training the x-vector on new data. As a point of reference, Kaldi's recipe yields ~**7–8% DER** on CALLHOME (oracle #speakers, 8kHz) and ~19–22% DER on DIHARD challenges (depending on clustering params). It set the *baseline* for diarization for years. However, it struggles with overlapping speech (which newer neural methods handle). *Release:* 2018 (Kaldi v5.0). Model size ~4M (TDNN). Still useful for fast, low-resource scenarios.

ASR Models Sorted by Release Date (newest → oldest):

Model	Release	Supported Languages	WER (ASR)	Size (params)
MMS (Meta)	2023	~1100 languages (extremely multilingual) [1]	~50% of Whisper's WER [2] (54-lang avg)	~1B

Model	Release	Supported Languages	WER (ASR)	Size (params)
OpenAI Whisper	2022	98 languages (multilingual) [3]	~9.9% (En, test avg) [4]	39M–1.55B
AI4Bharat IndicWav2Vec	2022	40 Indic languages [8]	16.6% (bn; 13.6% w/ LM) [10]	~317M
WavLM (Microsoft)	2021	English (pretrain)	1.8% / 3.2% (LS test) [12]	~300M
XLS-R (Meta)	2021	128 languages [15]	– (14–34% better vs XLSR) [17]	300M–2B
Wav2Vec 2.0 / XLSR-53	2020	53 languages (XLSR-53) [20]	1.8% / 3.3% (LS) [19]	~300M
QuartzNet (NVIDIA)	2019–20	English	3.2% / 7.5% (LS+LM) [25]	~18M
Jasper (NVIDIA)	2019	English	<3% (LS+LM) [28]	~332M
Mozilla DeepSpeech	2017–20	English	~7% (LS clean) [30]	~47M

※ LS = LibriSpeech. "En" = English. WER on LibriSpeech test-clean/test-other (unless noted).

Speaker Diarization Models Sorted by Release Date:

Model	Release	Approach	DER (benchmark)	Size
NVIDIA Sortformer	2023	End-to-end Transformer	8.5% (CALLHOME 3-spk) [35] ; 14.8% (DIHARD) [34]	~115M
NVIDIA MSDD (NeMo)	2021–22	Hybrid (embeddings + transf.)	8.2% (CallHome 2-spk) [36]	–
Pyannote toolkit	2020→2023	Pipeline (CNN + clustering)	17.0% (AMI) [46] ; 26.7% (CALLHOME) [47]	~20–30M
EEND (BLSTM)	2019	End-to-end (PIT BLSTM)	~20% (CALLHOME 2-spk) [42]	~7M
UIS-RNN (Google)	2019	Learned clustering (RNN)	~12% (CALLHOME 2-spk)	<1M
Kaldi x-vector	2018	Embedding + clustering	7–8% (CALLHOME 2-spk)	~4M

※ DER = Diarization Error Rate (lower is better). Results may vary across eval sets and conditions (numbers here give a sense of performance on common benchmarks as cited).

Sources: The information above is drawn from research papers and open-source documentation for each model, including WER/DER metrics on standard benchmarks. Key references: Wav2Vec 2.0 [19], XLSR [16], Whisper [4], IndicWav2Vec [10], Jasper/QuartzNet [28] [25], DeepSpeech [30], Pyannote [38], EEND [42], Sortformer [34], and others as cited above. Each model's repository (GitHub or Hugging Face) and associated paper are also indicated inline.

---

[1] [2] Meta's Open-Source Massively Multilingual Speech AI Handles over 1,100 Languages - InfoQ
https://www.infoq.com/news/2023/06/meta-mms-speech-ai/

[3] Introducing Whisper - OpenAI
https://openai.com/index/whisper/

[4] [5] cdn.openai.com
https://cdn.openai.com/papers/whisper.pdf

[6] [7] [2507.01931] Adaptability of ASR Models on Low-Resource Language: A Comparative Study of Whisper and Wav2Vec-BERT on Bangla
https://ar5iv.labs.arxiv.org/html/2507.01931

[8] [9] [10] GitHub - AI4Bharat/IndicWav2Vec: Pretraining, fine-tuning and evaluation scripts for Indic-Wav2Vec2
https://github.com/AI4Bharat/IndicWav2Vec

[11] Bengali.AI Speech Recognition - Kaggle Solutions
https://kaggle.curtischong.me/competitions/Bengali.AI-Speech-Recognition

[12] [13] [14] arxiv.org
https://arxiv.org/pdf/2110.13900

[15] [16] [17] [18] [22] arxiv.org
https://arxiv.org/pdf/2111.09296

[19] wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations - DOKUMEN.PUB
https://dokumen.pub/wav2vec-20-a-framework-for-self-supervised-learning-of-speech-representations.html

[20] [21] [23] XLSR-53: Crosslingual Wav2vec 2.0 Model
https://www.emergentmind.com/topics/crosslingual-wav2vec-2-0-model-xlsr-53

[24] [25] [26] arxiv.org
https://arxiv.org/pdf/1910.10261

[27] [28] [29] NVIDIA Releases New ASR Model and Speech Toolkit at Interspeech 2019 | NVIDIA Technical Blog
https://developer.nvidia.com/blog/new-asr-model-speech-toolkit-interspeech2019/

[30] What is WER/CER of DeepSpeech v0.7.1 (or any other models) on ...
https://discourse.mozilla.org/t/what-is-wer-cer-of-deepspeech-v0-7-1-or-any-other-models-on-common-voice-english/65600

[31] mozilla/DeepSpeech - GitHub
https://github.com/mozilla/DeepSpeech

[32] [33] [34] [35] [36] [44] arxiv.org
https://www.arxiv.org/pdf/2409.06656v2

[37] [38] [39] [41] [46] [47] GitHub - pyannote/pyannote-audio: Neural building blocks for speaker diarization: speech activity detection, speaker change detection, overlapped speech detection, speaker embedding
https://github.com/pyannote/pyannote-audio

[40] Top 8 speaker diarization libraries and APIs in 2025
https://www.assemblyai.com/blog/top-speaker-diarization-libraries-and-apis

[42] [43] [45] End-to-End Neural Diarization (EEND)
https://www.emergentmind.com/topics/end-to-end-neural-diarization-eend