Audio-to-Text Conversion - Model layer

Hugging Face Model Zoo — What to use and when

CTC on raw waveform (self-supervised encoders)

- **What it is:** Encoder (wav2vec 2.0 / HuBERT / WavLM) + CTC head over characters/BPE.
- **Why it matters:** Strong with limited labels; simple, fast decoding; easy domain finetuning.
- **Checkpoints:**
  - `facebook/wav2vec2-base-960h` (EN)
  - `facebook/wav2vec2-large-960h-lv60-self` (EN, SSL-pretrained)
  - `facebook/wav2vec2-large-xlsr-53` (multilingual pretrain; many finetunes exist per language)
  - `facebook/hubert-large-ls960-ft` (EN)
  - `microsoft/wavlm-base-plus` (pretrained; add a CTC head)
- **Best for:** Read/clean speech, telephony after light domain adaptation, low-latency CTC decoding.

Seq2Seq on log-Mel (Whisper family)

- **What it is:** Encoder-decoder Transformer trained on huge weakly-labeled pairs; predicts text tokens (and timestamps).
- **Why it matters:** SOTA robustness, multilingual, works well zero-shot; timestamps & translation modes.
- **Checkpoints:**
  - `openai/whisper-tiny` / `…-base` / `…-small` / `…-medium` / `…-large-v2` / `…-large-v3`
  - Distilled: `distil-whisper/distil-small.en`, `…/distil-large-v2` (faster)
  - Many domain/language finetunes (e.g., `NbAiLab/whisper-small-nob` for Norwegian)
- **Best for:** Noisy, accented, long-form audio; multilingual; needs timestamps or translation.

Conformer / Transducer (streaming-friendly)

- **What it is:** Conformer (conv + attention) encoder with CTC/Transducer decoding; supports streaming.

- **Why it matters:** Low latency and strong WER, popular in production.
- **Checkpoints (HF hosted by frameworks):**
  - SpeechBrain Conformer+TransformerLM (e.g., `speechbrain/asr-conformer-transformerlm-librispeech`)
  - ESPnet Conformer (e.g., `espnet/kan-bayashi_librispeech_asr_train_asr_conformer_raw_bpe_sp_valid.acc.ave`)
  - Meta wav2vec2-Conformer finetunes: `facebook/wav2vec2-conformer-rel-pos-large-960h-ft`
- **Best for:** Online/streaming ASR, meetings/calls where latency is key.

Massively Multilingual (low-resource coverage)

- **What it is:** Models trained across 100+ languages; handles very low-resource locales.
- **Why it matters:** Coverage beats per-language models when data is scarce.
- **Checkpoints:**
  - `facebook/mms-1b-all` / `facebook/mms-300m` (MMS ASR)
  - Whisper multilingual sizes (above)
- **Best for:** Broad language support, cross-lingual transfer.

Domain-specific finetunes

- **What it is:** Finetunes on finance, medical, meetings, etc.
- **Why it matters:** Big gains on jargon, numerals, and acoustics.
- **Examples:** Earnings calls (`mozilla-foundation/earnings22` finetunes on Whisper); many language-specific wav2vec2 finetunes like `jonatasgrosman/wav2vec2-large-xlsr-53-xx-xx`.

Add-ons you'll likely want

- **Punctuation/casing restoration:** Run a lightweight seq2seq/transformer after ASR (search "punctuation restoration" models on HF).
- **CTC + LM decoding:** `pyctcdecode` + KenLM for numbers, acronyms, domain terms.
- **Diarization:** Pipeline ASR with pyannote models when multi-speaker timestamps matter.

Architectural innovations (why these work)

- **Self-supervised pretraining on raw audio (wav2vec2, HuBERT, WavLM):** contrastive/masked prediction creates strong acoustic representations from hours of unlabeled speech; CTC head fine-tunes with little labeled data.

- **Conformer blocks:** marry local conv (phones/formants) with global attention (long context), improving WER and efficiency, and enabling streaming variants.
- **Transducer decoding (RNN-T/Conformer-T):** alignment-free, low-latency decoding better suited to streaming than full attention decoders.
- **Seq2Seq (Whisper):** encoder-decoder Transformer trained on massive weakly-labeled pairs; built-in language/task tokens, timestamp prediction, and robustness to noise/domain shift.
- **Multilingual subword vocabularies:** shared tokenizers/BPE that generalize across languages and handle code-switching.
- **SpecAugment & speed perturbation baked in training:** regularization that survives domain shift.
