

NAME

cmtranscribe - speech to text with a local Parakeet TDT model

SYNOPSIS

cmtranscribe [*OPTION*]... *MODEL AUDIO*...

DESCRIPTION

cmtranscribe loads *MODEL* and writes a transcript of each *AUDIO*. It runs NVIDIA's Parakeet TDT 0.6B v3 architecture (FastConformer encoder, token-and-duration transducer decoder) on the CPU, with punctuation and casing as the model produces them.

MODEL is Fermion Research's Phonon-2 (**FermionResearch/Phonon-2** on Hugging Face) in any of three forms: the snapshot directory the download created, the **.tar.zst** archive in it, or a directory with the archive's unpacked **config.json** and **model.fermion**. It can also be a NeMo-named Parakeet TDT GGUF (architecture **asr**), such as **parakeet-tdt-0.6b-v3.q8_0.gguf** in **nvidia/parakeet-tdt-0.6b-v3**, which covers 25 European languages; or a directory holding one.

AUDIO is a WAV file (PCM 8, 16, 24 or 32 bit, or 32 or 64 bit float), or - for standard input. Channels are averaged; other sample rates are resampled to 16 kHz. Other formats can be piped through ffmpeg:

```
ffmpeg -i talk.mp3 -f wav - | cmtranscribe MODEL -
```

OPTIONS

-f, --format *FMT*

text (default): the transcript, prefixed with the file name when there are several. **json** adds the duration and each word with start and end times in seconds. **srt** and **vtt** write subtitles; a cue ends after a sentence, before a pause over a second, or at 7 seconds or 84 characters. Subtitles take one *AUDIO*.

-o, --output *PATH*

Write to *PATH* instead of standard output.

--chunk *SECONDS*

Transcribe audio longer than this in pieces, each ending at the quietest 80 ms in its last quarter. Attention costs grow with the square of the length, so one piece of several minutes is slow. 0 transcribes in one piece. Default: 30.

-v, --verbose

Report the load time, and per file the audio length, the time taken and the speed as a multiple of real time, on standard error.

-h, --help

Show the options.

ENVIRONMENT**CMGENERATE_THREADS**

Worker threads.

CMGENERATE_KERNELS

float (or **avx2**, **scalar**) keeps the f32 products. Otherwise, on a CPU with AVX-512 VNNI, the encoder multiplies 8-bit weights by activations quantised to 8 bits per frame, about 1.6 times faster; transcripts can differ where two words score almost the same.

EXAMPLES

```
cmtranscribe ~/.cache/huggingface/hub/models--FermionResearch--Phonon-2/snapshots/* talk.wav
```

```
cmtranscribe -f srt -o talk.srt phonon-2 talk.wav
```

SEE ALSO

cmgenerate(1), **ffmpeg**(1)