NAME

cmtranscribe − speech to text with a local Parakeet model

SYNOPSIS

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

DESCRIPTION

cmtranscribe loads *MODEL* and writes a transcript of each *AUDIO*. It runs NVIDIA's Parakeet architecture (FastConformer encoder, 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.

MODEL can also be a NeMo **.nemo** checkpoint, or a directory holding one, of a cache-aware streaming RNNT model such as **nvidia/multitalker−parakeet−streaming−0.6b−v1** (English). Such a model has causal convolutions and attends to whole chunks of 80 ms frames, so it can transcribe as audio arrives; see **−−live** and **−−lookahead**. Its word times are when the model emitted the word: on LibriSpeech, a median 0.56 s (**−−lookahead** 13) to 0.72 s (0) after Parakeet TDT's.

AUDIO is a WAV file (PCM 8, 16, 24 or 32 bit, or 32 or 64 bit float), any file symphonia decodes (FLAC, MP3, Ogg Vorbis, AAC or ALAC in MP4/M4A, AIFF, CAF, MKV/WebM; not Opus), or − for standard input. Channels are averaged; other sample rates are resampled to 16 kHz. Opus (e.g. a browser recording in WebM or Ogg) can be piped through ffmpeg:

ffmpeg −i talk.webm −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. A streaming model reads the pieces as one stream; the result equals one pass.

−−live

Streaming models only: read *AUDIO* as it arrives and write the text of each chunk as soon as the chunk is complete. *AUDIO* is 16-bit mono PCM at the model's rate (16 kHz), raw or in a WAV header; one *AUDIO* and the text format only.

−−diarize *DIAR*

Streaming models only: tell up to four speakers apart with a streaming Sortformer diarizer, *DIAR* being its **.nemo** or the directory holding it, such as **nvidia/diar_streaming_sortformer_4spk−v2.1**. As NeMo's multitalker pipeline does, each 1.12 s chunk is diarized first; every speaker active in the last two chunks then has the chunk transcribed by their own instance of the model, with the frames where they are silent zeroed. Speakers are numbered in the order they first speak. Text output has a line per run of one speaker's words (split at pauses over 1 s), *m:ss*-*m:ss* **Speaker** *N***:** then the words. Runs that overlap in time form a block of consecutive lines; blocks are separated by an empty line. **json** adds **turns** (these lines, with their **block**) and a **speaker** per word; **srt** and **vtt** start each cue

with the speaker. With **−−live**, a line starts at each change of speaker. A speaker's instance only runs while they speak, so a period closing a turn can arrive with their next turn; live output drops it there.

−−speaker−kernels
With **−−diarize**: give each instance the whole chunk and steer it with the model's speaker kernels, fed the speaker's probabilities and the other speakers' activity (NeMo's **masked_asr** false, self-speaker adaptation), instead of zeroing the frames where the speaker is silent. Overlapped speech then reaches every instance.

−−lookahead *FRAMES*
Streaming models only: the right attention context, one of those the model was trained with (13, 6, 1 or 0 for the multitalker model). A chunk is *FRAMES* + 1 frames of 80 ms, which sets the latency: 1.12 s for 13, 0.08 s for 0. Default: the first in the model's config.

−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; **−−threads** *N* overrides it.

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
arecord −f S16_LE −r 16000 −c 1 −t raw | cmtranscribe −−live multitalker−parakeet−streaming−0.6b−v1.nemo −
cmtranscribe −−diarize diar_streaming_sortformer_4spk−v2.1.nemo multitalker−parakeet−streaming−0.6b−v1.nemo meeting.wa

SEE ALSO

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