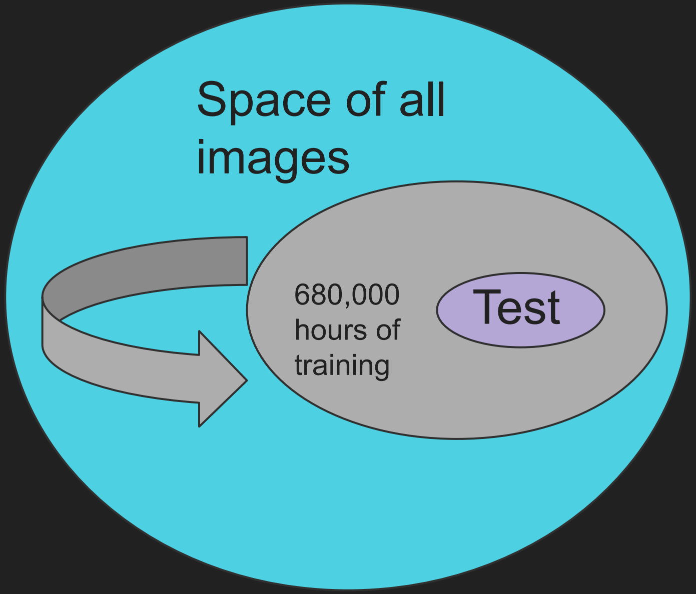
Speech-to-Text: Whisper

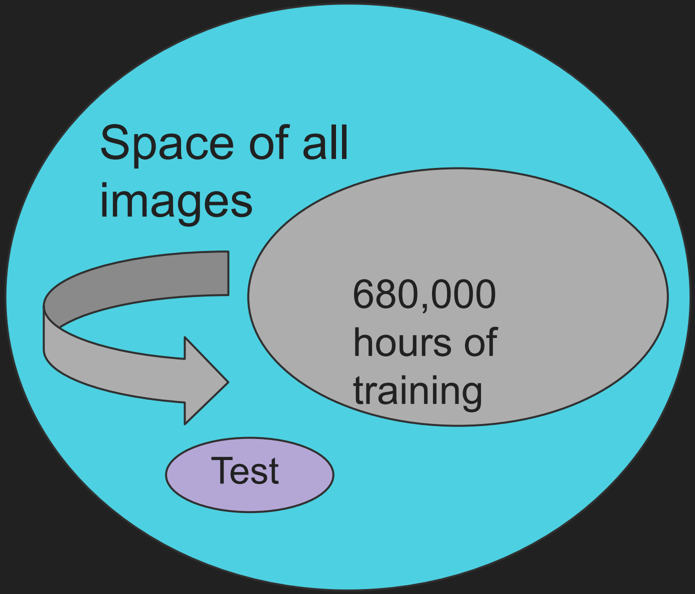
Varshini Narayanan

Introduction

- WHISPER OPEN AI - ASR(Automatic Speech Recognition)
- 680k hours of multilingual labelled data (librispeech)



‘In’ distribution

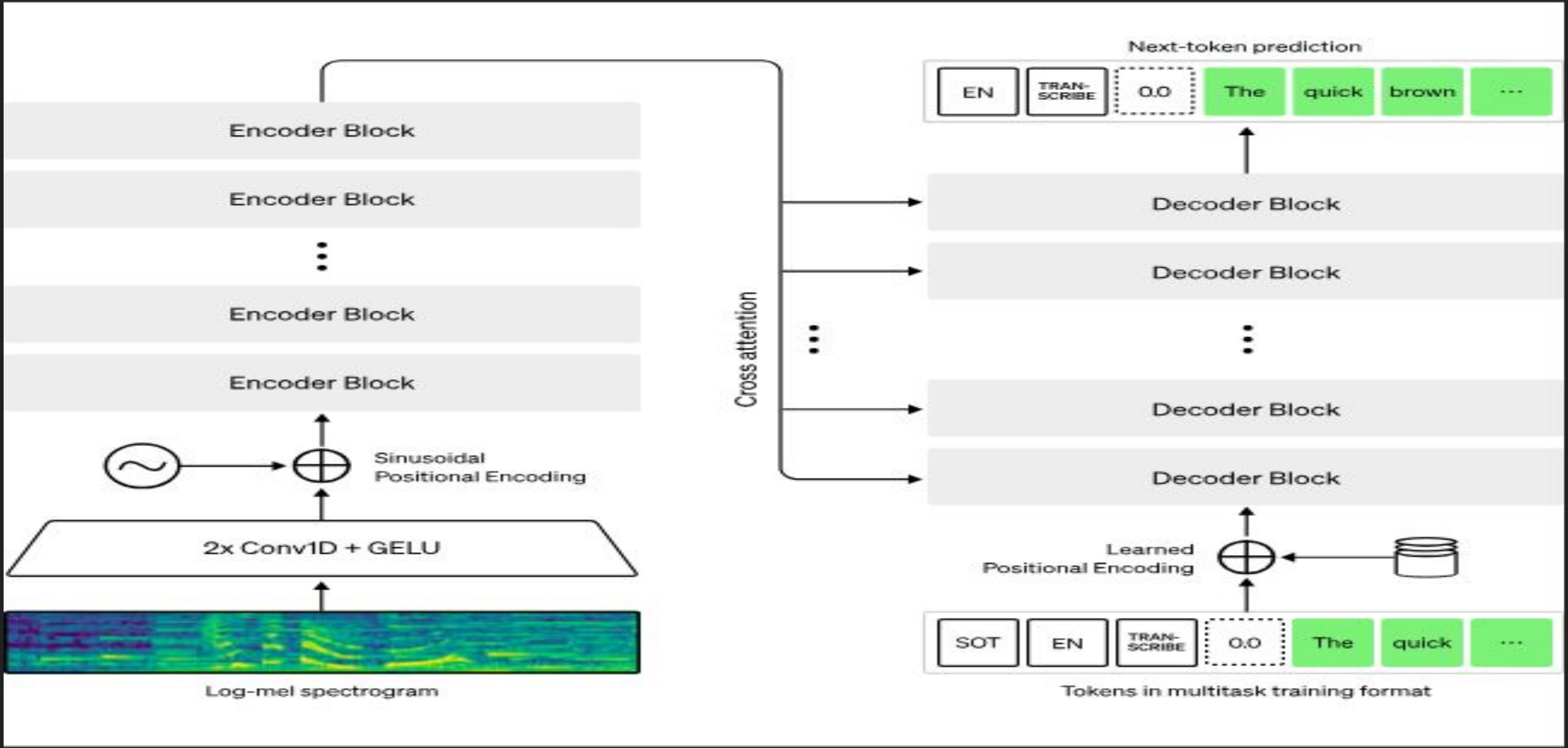


‘Out’ distribution

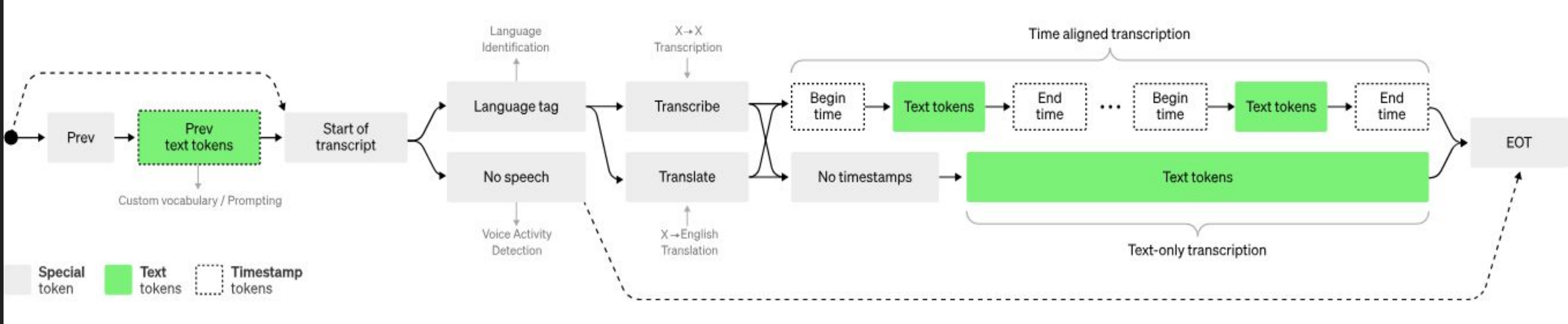
Whisper large-v3

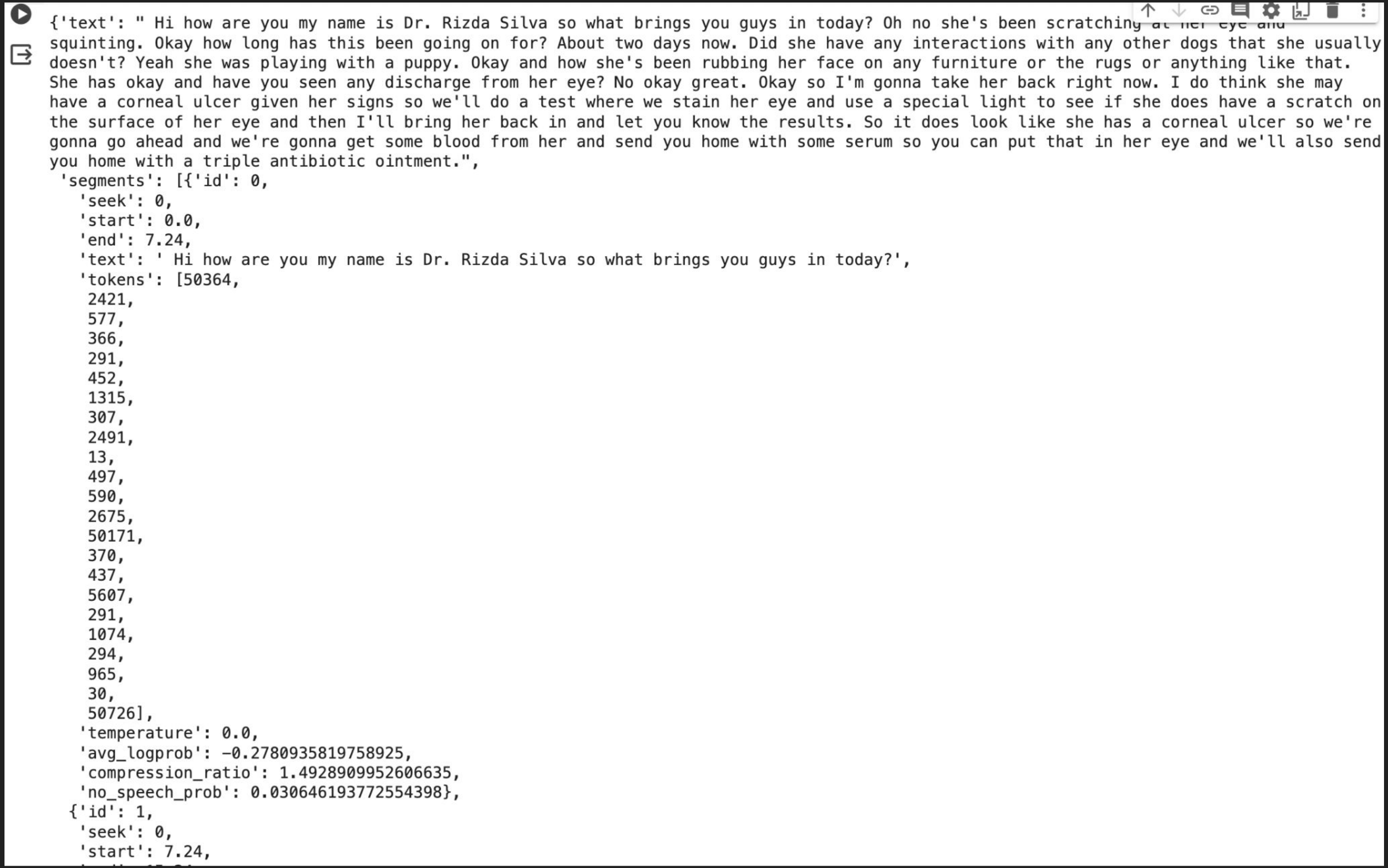
- Transformer (encoder-decoder), sequence-to-sequence model (99 languages)
- The input uses 128 Mel frequency bins instead of 80
- A new language token for Cantonese
- 1 million hours of weakly labeled audio
- 4 million hours of pseudolabeled audio from large-v2
- epochs=2.0 more
- English texts (only trained for transcription purposes)
- Multilingual texts (were trained for both transcription and translation)
- 32 Transformer layers with a dimension of 1280
- Activation function: GELU (Gaussian Error Linear Unit)
- Optimizer: Adam
- Evaluation Metric: WER (word error rate)

Architecture

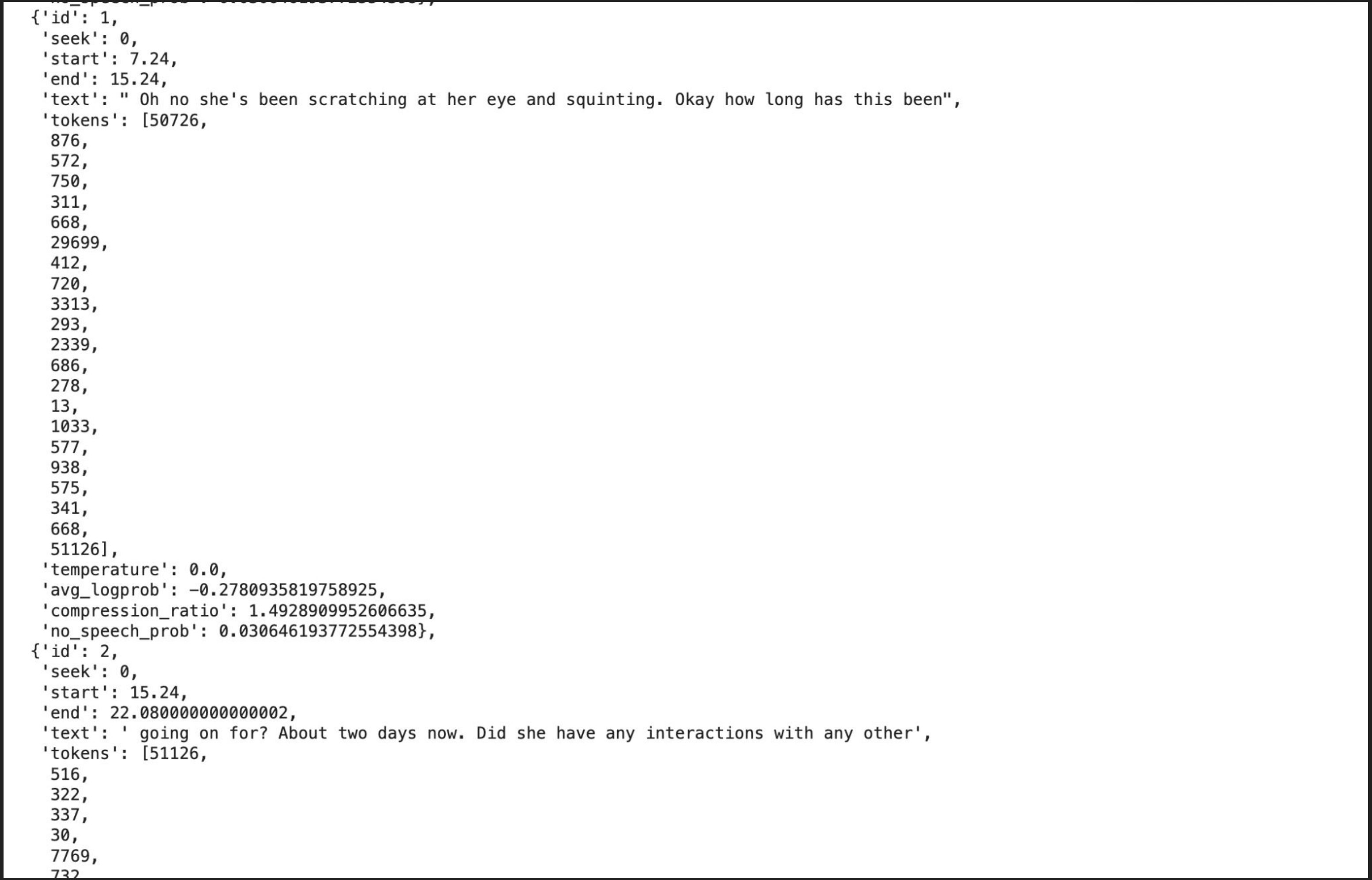


Token Architecture:



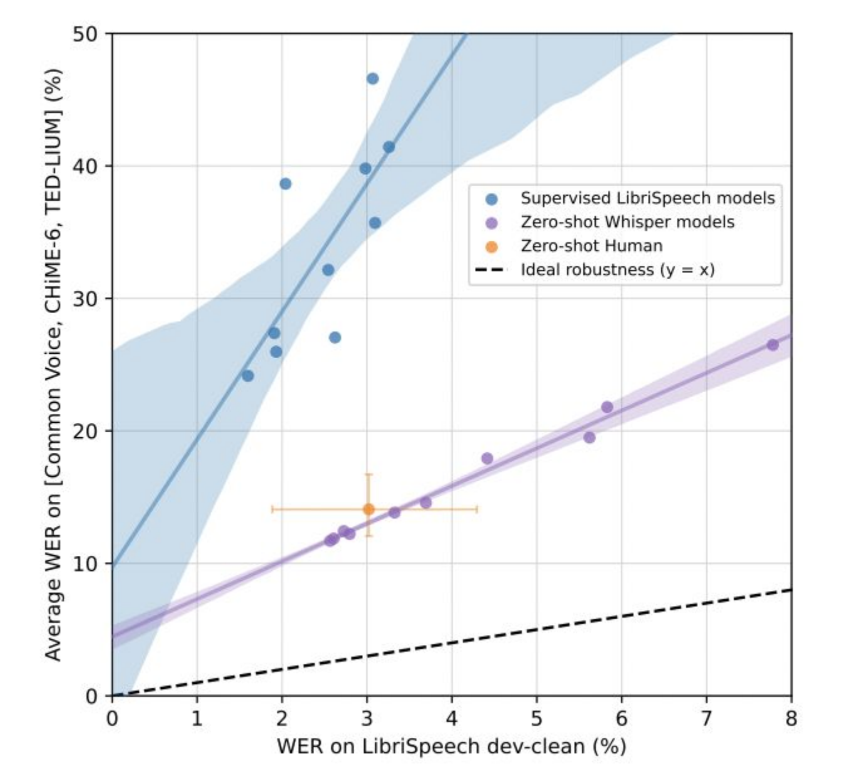


```
{'text': " Hi how are you my name is Dr. Rizda Silva so what brings you guys in today? Oh no she's been scratching at her eye and
squinting. Okay how long has this been going on for? About two days now. Did she have any interactions with any other dogs that she usually
doesn't? Yeah she was playing with a puppy. Okay and how she's been rubbing her face on any furniture or the rugs or anything like that.
She has okay and have you seen any discharge from her eye? No okay great. Okay so I'm gonna take her back right now. I do think she may
have a corneal ulcer given her signs so we'll do a test where we stain her eye and use a special light to see if she does have a scratch on
the surface of her eye and then I'll bring her back in and let you know the results. So it does look like she has a corneal ulcer so we're
gonna go ahead and we're gonna get some blood from her and send you home with some serum so you can put that in her eye and we'll also send
you home with a triple antibiotic ointment.",
 'segments': [{'id': 0,
   'seek': 0,
   'start': 0.0,
   'end': 7.24,
   'text': ' Hi how are you my name is Dr. Rizda Silva so what brings you guys in today?',
   'tokens': [50364,
    2421,
    577,
    366,
    291,
    452,
    1315,
    307,
    2491,
    13,
    497,
    590,
    2675,
    50171,
    370,
    437,
    5607,
    291,
    1074,
    294,
    965,
    30,
    50726],
   'temperature': 0.0,
   'avg_logprob': -0.2780935819758925,
   'compression_ratio': 1.4928909952606635,
   'no_speech_prob': 0.030646193772554398},
  {'id': 1,
   'seek': 0,
   'start': 7.24,
```



```
 'no_speech_prob': [illegible]},
{'id': 1,
 'seek': 0,
 'start': 7.24,
 'end': 15.24,
 'text': " Oh no she's been scratching at her eye and squinting. Okay how long has this been",
 'tokens': [50726,
  876,
  572,
  750,
  311,
  668,
  29699,
  412,
  720,
  3313,
  293,
  2339,
  686,
  278,
  13,
  1033,
  577,
  938,
  575,
  341,
  668,
  51126],
 'temperature': 0.0,
 'avg_logprob': -0.2780935819758925,
 'compression_ratio': 1.4928909952606635,
 'no_speech_prob': 0.030646193772554398},
{'id': 2,
 'seek': 0,
 'start': 15.24,
 'end': 22.080000000000002,
 'text': ' going on for? About two days now. Did she have any interactions with any other',
 'tokens': [51126,
  516,
  322,
  337,
  30,
  7769,
  732
```

Evaluation Metric: WER (word error rate)

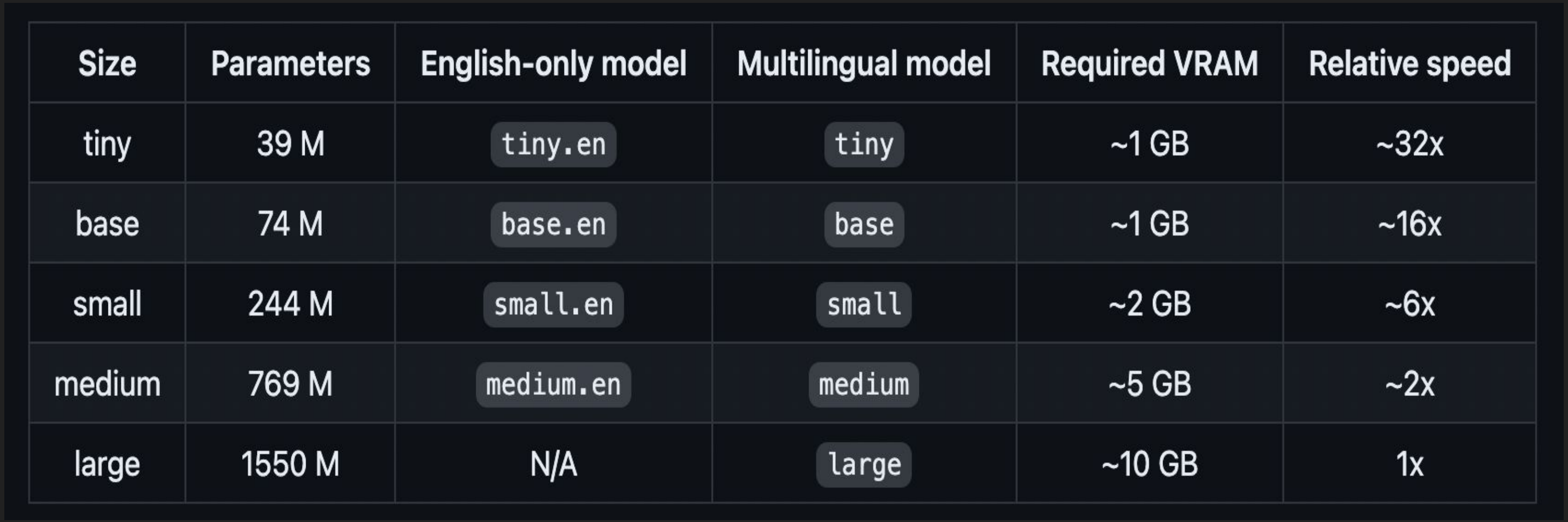


- Other librispeech models (resnet) is far away from the human predicted point.
- Whisper is closer to the zero-shot human predicted point (more robust and is almost equal to human prediction)

Use cases:

- Multilingual speech transcription
- To-English speech translation
- Language identification
- Voice Activity Detection (less accurate)

Different Configurations:



Size	Parameters	English-only model	Multilingual model	Required VRAM	Relative speed
tiny	39 M	`tiny.en`	`tiny`	~1 GB	~32x
base	74 M	`base.en`	`base`	~1 GB	~16x
small	244 M	`small.en`	`small`	~2 GB	~6x
medium	769 M	`medium.en`	`medium`	~5 GB	~2x
large	1550 M	N/A	`large`	~10 GB	1x

Limitations:

- Inaccurate timestamps
- Low Performance
- No built-in voice Diarization
- No real-time transcription
- Pure PyTorch inference

whisperX

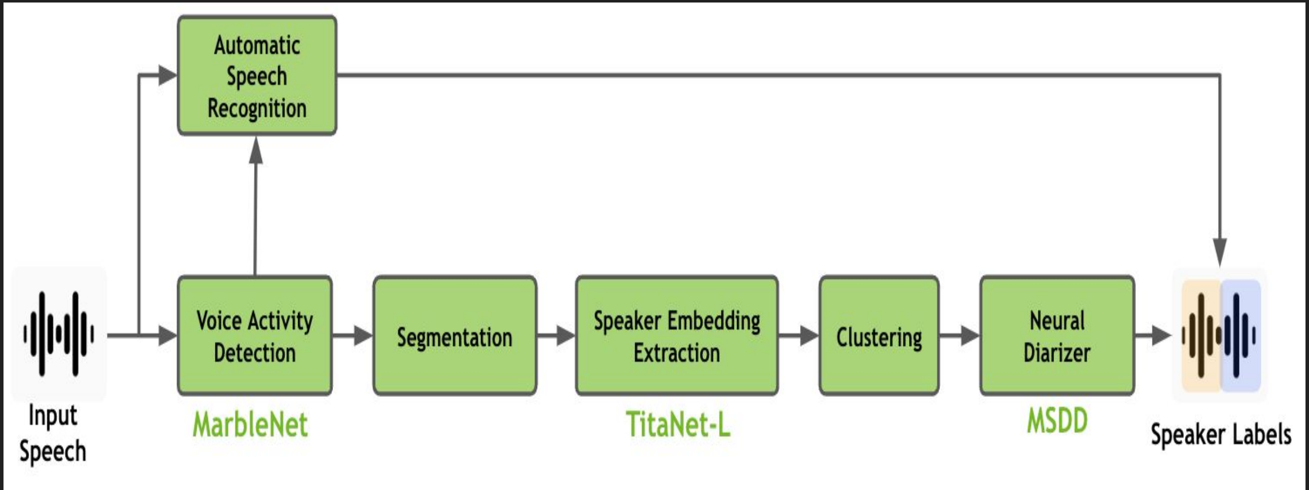
- 70x realtime with large-v2
- Word level timestamps
- Speaker diarization

VAD ?

- Voice activity detection (to identify/detect human speech in an audio)
- Very effective for speech transcription purposes

Speech Embedding?

- Conversion of raw input -> numerical array representation (vector) with a certain dimension.
- Used for speech diarizations where the difference in speaker is identified with separating different segments of the labels.

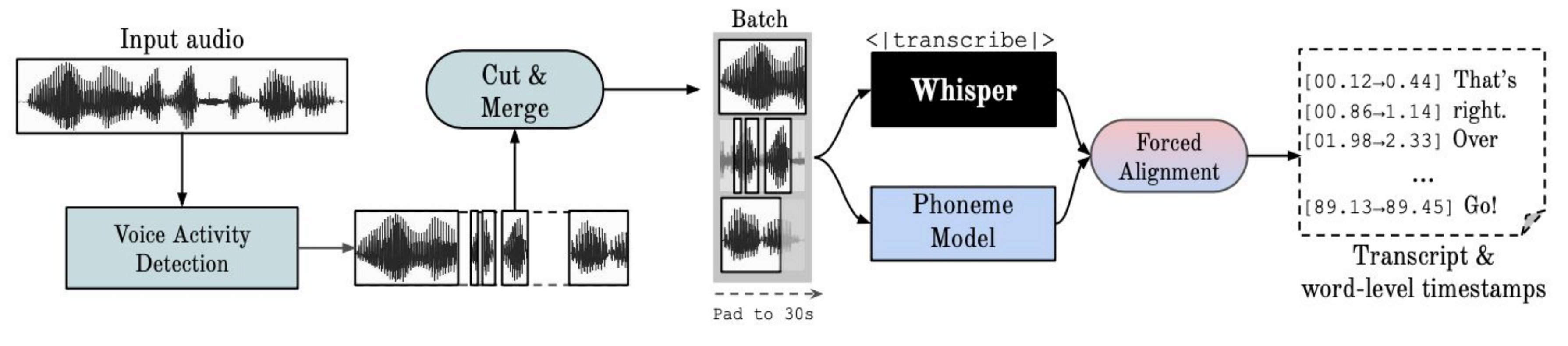


WhisperX + MarbleNet(VAD) + TitaNet (Speech Embedding): NeMo compatability

Steps:

- Vocals are extracted from the audio (speaker embedding accuracy increases)
- Transcription generated using Whisper (timestamp tokens)
- Timestamps collected from whisperX = timestamp tokens of whisper
- MarbleNet (segmentation) (VAD) - to exclude silence
- TitaNet is then used to extract speaker embeddings to identify the speaker for each segment
- Realigned using punctuation models to compensate for minor time shifts.

WhisperX



Input audio
Voice Activity Detection
Cut & Merge
Batch
Pad to 30s
<|transcribe|>
Whisper
Phoneme Model
Forced Alignment
[00.12→0.44] That's
[00.86→1.14] right.
[01.98→2.33] Over
...
[89.13→89.45] Go!
Transcript & word-level timestamps

MarbleNet

- Multinodal MarbleNet model
- It is a 1D resnet model that uses batch normalizations, ReLu Activation Function and dropout layers.
- Used for VAD (Voice Activity Detection)
- VAD ensures whether a speaker is present or not in the audio

TitaNet

- Depth wise 1D conv model that helps in speech embedding.
- Speech embedding: raw speech signals -> fixed-dimensional numerical representations (vectors)
- Only uses the important part of the speech

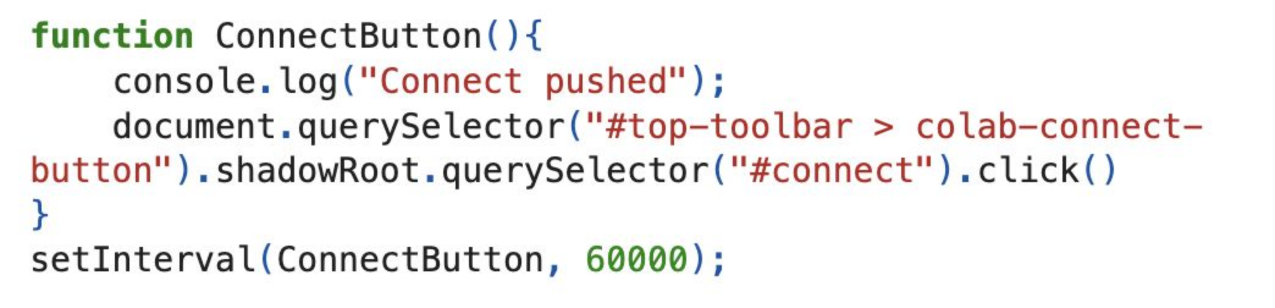
Common Errors

Out-of-Memory:

- Use T4 GPU with correct CUDA version (reduce no of steps in the model)

ImportError:

- Importing the correct libraries



```
function ConnectButton(){
    console.log("Connect pushed");
    document.querySelector("#top-toolbar > colab-connect-button").shadowRoot.querySelector("#connect").click()
}
setInterval(ConnectButton, 60000);
```

Future Aspects

What are the issues while looking more real-time?

- Overlapping of speech
- Extra language model like wave2vec2 is needed

Overcome:

- Apple MLX (fine tuning for Llama factory - visually based)
- Ferret

(No of parameters: 7B/13B parameters)