Whisper Live installation guide

> **Note:** it is preferable to use a machine with NVidia GPU in order to use the most accurate whisper models. Else on CPU most probably you will be limited to 'base' or 'small' models.

1. Install Docker in it is not already installed. For Windows install Docker Desktop from here: https://www.docker.com/products/docker-desktop/
2. If you have NVidia GPU make sure you have installed the latest drivers from NVidia
3. Install and run Whisper Live:

    - No GPU:

    Powershell:

```
docker run `
 --name whisper-live `
 --restart=unless-stopped `
 -e WHISPERLIVE_API_KEY=mysecretkey `
 -e WHISPERLIVE_MODEL=base `
 -e WHISPERLIVE_MAX_CLIENTS=4 `
 -e WHISPERLIVE_MAX_CONNECTION_TIME=300 `
 -e WHISPERLIVE_USE_VAD=true `
 -v whisper-live-data:/var/lib/whisper-live `
 -p 9090:9090 `
 -p 8000:8000 `
 -d hwdsl2/whisper-live-server
```

    Bash:

```
docker run \
 --name whisper-live \
 --restart=unless-stopped \
 -e WHISPERLIVE_API_KEY=mysecretkey \
 -e WHISPERLIVE_MODEL=base \
 -e WHISPERLIVE_MAX_CLIENTS=4 \
 -e WHISPERLIVE_MAX_CONNECTION_TIME=300 \
 -e WHISPERLIVE_USE_VAD=true \
 -v whisper-live-data:/var/lib/whisper-live \
```

```
 -p 9090:9090 \
 -p 8000:8000 \
 -d hwdsl2/whisper-live-server
```

- NVIDIA GPU:

Powershell:

```
 docker run `
 --name whisper-live `
 --gpus=all `
 --restart=unless-stopped `
 -e WHISPERLIVE_API_KEY=mysecretkey `
 -e WHISPERLIVE_MODEL=base `
 -e WHISPERLIVE_MAX_CLIENTS=4 `
 -e WHISPERLIVE_MAX_CONNECTION_TIME=300 `
 -e WHISPERLIVE_USE_VAD=true `
 -v whisper-live-data:/var/lib/whisper-live `
 -p 9090:9090 `
 -p 8000:8000 `
 -d hwdsl2/whisper-live-server:cuda
```

Bash:

```
sudo docker run \
 --name whisper-live \
 --gpus=all \
 --restart=unless-stopped \
 -e WHISPERLIVE_API_KEY=mysecretkey \
 -e WHISPERLIVE_MODEL=base \
 -e WHISPERLIVE_MAX_CLIENTS=4 \
 -e WHISPERLIVE_MAX_CONNECTION_TIME=300 \
 -e WHISPERLIVE_USE_VAD=true \
 -v whisper-live-data:/var/lib/whisper-live \
 -p 9090:9090 \
 -p 8000:8000 \
 -d hwdsl2/whisper-live-server:cuda
```

Change the values of the environment as per your needs, for example **WHISPERLIVE_MAX_CONNECTION_TIME=300** allow sessions up to 300 seconds (5 minutes), so you may want this to be **WHISPERLIVE_MAX_CONNECTION_TIME=3600** for example.

Important ones are the model and API key. For GPU the good results are achieved with **large-v3-turbo**.

For RHEL default `podman` (official RHEL docker replacement) for NVidia GPU instead of `--gpus=all` use `--device nvidia.com/gpu=...` Get the full name of the device from output of `nvidia-ctk cdi list`.

Available models:

Model	Disk	RAM (approx)	Notes
`tiny`	~75 MB	~250 MB	Fastest; lower accuracy
`tiny.en`	~75 MB	~250 MB	English-only
`base`	~145 MB	~700 MB	Good balance — **default**
`base.en`	~145 MB	~700 MB	English-only
`small`	~465 MB	~1.5 GB	Better accuracy
`small.en`	~465 MB	~1.5 GB	English-only
`medium`	~1.5 GB	~5 GB	High accuracy
`medium.en`	~1.5 GB	~5 GB	English-only
`large-v1`	~3 GB	~10 GB	Older large model
`large-v2`	~3 GB	~10 GB	Very high accuracy
`large-v3`	~3 GB	~10 GB	Best accuracy
`large-v3-turbo`	~1.6 GB	~6 GB	Fast + high accuracy
`turbo`	~1.6 GB	~6 GB	Alias for `large-v3-turbo`

If you need to start the container with another parameters do:

```
docker rm -f whisper-live
```

and then start it again with the new parameters. Note that the downloaded models will be preserver in **whisper-live-data** volume.

In SubtitleNEXT use the following URL: ws://127.0.0.1:9090/ (if it is on different machine use its IP instead) and for API key the value you have set to **WHISPERLIVE_API_KEY** ('*mysecretkey*' in the examples above).

> **Note:** On first use the configured model will be downloaded, so the API will not respond until it finishes. Wait either until it start return results or if connection is closed try again later.

Revision #6
Created 5 October 2026 09:33:23 by Pierre
Updated 20 November 2026 12:02:16 by Pierre