Kokoro

Overview

Kokoro is a high-quality, locally-hosted text-to-speech (TTS) service. It exposes an OpenAI-compatible TTS API, making it a drop-in replacement for external TTS services in any application that supports the `/v1/audio/speech` endpoint.

It runs the Kokoro TTS model via the `kokoro-fastapi` server with full RTX 5080 GPU acceleration.

Access

Type	URL	Notes
Internal	`https://kokoro.home.local`	LAN access via Step-CA TLS

No external route — internal service only.

Configuration

Image: `ghcr.io/remsky/kokoro-fastapi-gpu:latest-cu128` **Compose project:** `ai-stack` **Runtime:** `nvidia` **CUDA Version:** 12.8 (image) / 12.9 (host driver, backward compatible)

Ports

Port	Protocol	Purpose
`8880`	TCP	Kokoro FastAPI TTS API

Traefik Labels

```
traefik.enable: "true"
traefik.docker.network: traefik-net
traefik.http.routers.kokoro.rule: Host(`kokoro.home.local`)
traefik.http.routers.kokoro.entrypoints: websecure
traefik.http.routers.kokoro.tls.certresolver: step-ca
traefik.http.services.kokoro.loadbalancer.server.port: 8880
```

GPU Acceleration

```
runtime: nvidia
environment:
  - NVIDIA_VISIBLE_DEVICES=all
  - DEVICE=gpu
  - USE_GPU=true
```

The image tag `latest-cu128` targets CUDA 12.8. The host driver (580.159.03) is fully forward-compatible with this image.

Volumes / Bind Mounts

Host Path	Container Path	Purpose
`/home/jeeves/docker/kokoro`	`/app/data`	Voice models and cached data

Networks

Network	Purpose
`ai-stack_ai-internal`	Internal access from Open Web UI and N8N
`traefik-net`	Exposes API via Traefik

API Usage

Kokoro exposes an OpenAI-compatible TTS endpoint:

```
curl -X POST https://kokoro.home.local/v1/audio/speech \
  -H "Content-Type: application/json" \
  -d '{
    "model": "kokoro",
    "input": "Hello from the homelab.",
    "voice": "af_sky",
    "response_format": "mp3"
  }' --output speech.mp3
```

Available voices and model details are listed at `https://kokoro.home.local/docs` (Swagger UI).

Dependencies

- NVIDIA container runtime with RTX 5080 access

Notes / Gotchas

- The `latest-cu128` tag pulls the latest build compiled against CUDA 12.8. NVIDIA driver 580.x supports CUDA 12.9 on the host, which is backward-compatible.
- Voice model files are cached in the `/app/data` bind mount on first use — initial synthesis requests may be slower while models are downloaded.
- Kokoro is OpenAI API-compatible, meaning Open Web UI can be configured to use it as its TTS provider via the admin settings.

Last Updated: 2026-06-16

Revision #3
Created 2026-06-17 04:43:12 UTC by Admin
Updated 2026-06-17 16:52:13 UTC by Admin