KISSKI
KI-Servicezentrum für sensible und kritische Infrastrukturen

Narges Lux

Enhancing Accessibility Through Audio-Visual Transcription and Translation Services

January 28, 2026

Table of contents

1 Introduction

2 Voice AI-Use Cases

3 Status

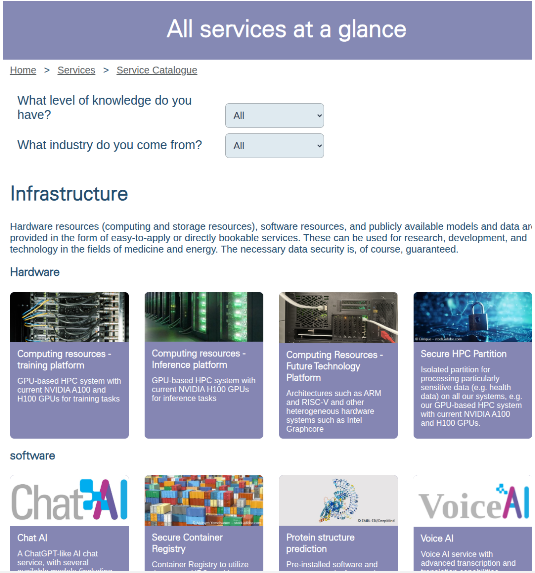
4 Challenges

5 Roadmap

AI Service Centre KISSKI

KISSKI: AI Service Centre for Sensitive and Critical Infrastructures

- **Goal**: Establish an AI Service Centre that meets the requirements of critical infrastructures, provides services for pilot studies throughout Germany, and continues operations after the funding period.
- **Requirements**: Ensuring security, privacy, and reliability for critical infrastructures.
- **Services**: Providing accessible AI infrastructure and expertise, open to researchers and industry, especially Start-Ups & SMEs.
- **Research**: Conducting research to further improve the services in terms of scalability, data management, and portability.



All services at a glance

Home > Services > Service Catalogue

What level of knowledge do you have? All

What industry do you come from? All

Infrastructure

Hardware resources (computing and storage resources), software resources, and publicly available models and data ar provided in the form of easy-to-apply or directly bookable services. These can be used for research, development, and technology in the fields of medicine and energy. The necessary data security is, of course, guaranteed.

Hardware

Computing resources - training platform
GPU-based HPC system with current NVIDIA A100 and H100 GPUs for training tasks

Computing resources - Inference platform
GPU-based HPC system with current NVIDIA H100 GPUs for inference tasks

Computing Resources - Future Technology Platform
Architectures such as ARM and RISC-V and other heterogeneous hardware systems such as Intel Graphcore

Secure HPC Partition
Isolated partition for processing particularly sensitive data (e.g. health data) on all our systems, e.g. our GPU-based HPC system with current NVIDIA A100 and H100 GPUs.

© Enrique - stock.adobe.com

software

Chat AI
A ChatGPT-like AI chat service, with several

Secure Container Registry
Container Registry to utilize

© Maksym Yemelyanov – stock.adobe.com

Protein structure prediction
Pre-installed software and

© EMBL-EBI/DeepMind

Voice AI
Voice AI service with advanced transcription and

Outline

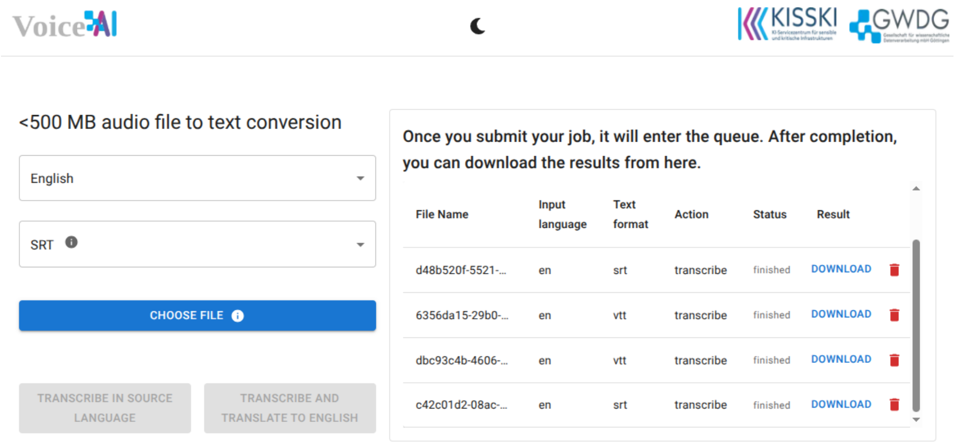
1 Introduction

2 Voice AI-Use Cases

3 Status

4 Challenges

5 Roadmap



<500 MB audio file to text conversion

English

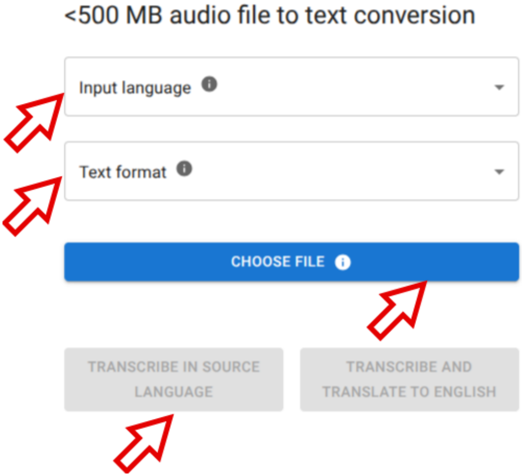
SRT

CHOOSE FILE
TRANSCRIBE IN SOURCE LANGUAGE
TRANSCRIBE AND TRANSLATE TO ENGLISH

Once you submit your job, it will enter the queue. After completion, you can download the results from here.

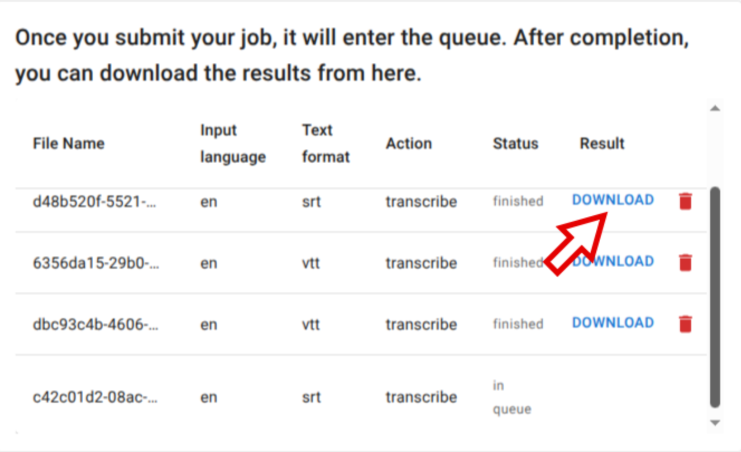
File Name	Input language	Text format	Action	Status	Result	
d48b520f-5521-...	en	srt	transcribe	finished	DOWNLOAD	🗑
6356da15-29b0-...	en	vtt	transcribe	finished	DOWNLOAD	🗑
dbc93c4b-4606-...	en	vtt	transcribe	finished	DOWNLOAD	🗑
c42c01d2-08ac-...	en	srt	transcribe	finished	DOWNLOAD	🗑

- **Audio file transcription**:



<500 MB audio file to text conversion

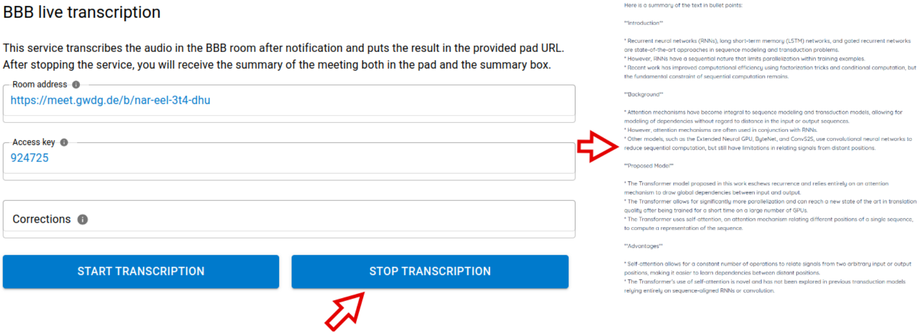
Input language
Text format
CHOOSE FILE
TRANSCRIBE IN SOURCE LANGUAGE
TRANSCRIBE AND TRANSLATE TO ENGLISH



Once you submit your job, it will enter the queue. After completion, you can download the results from here.

File Name	Input language	Text format	Action	Status	Result	
d48b520f-5521-...	en	srt	transcribe	finished	DOWNLOAD	🗑
6356da15-29b0-...	en	vtt	transcribe	finished	DOWNLOAD	🗑
dbc93c4b-4606-...	en	vtt	transcribe	finished	DOWNLOAD	🗑
c42c01d2-08ac-...	en	srt	transcribe	in queue		

- **Live BBB transcription**:



BBB live transcription

This service transcribes the audio in the BBB room after notification and puts the result in the provided pad URL. After stopping the service, you will receive the summary of the meeting both in the pad and the summary box.

Room address ⓘ
https://meet.gwdg.de/b/nar-eel-3t4-dhu

Access key ⓘ
924725

Corrections ⓘ

Here is a summary of the text in bullet points:

Introduction

* Recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and gated recurrent networks are state-of-the-art approaches in sequence modeling and transduction problems.
* However, RNNs have a sequential nature that limits parallelization within training examples.
* Recent work has improved computational efficiency using factorization tricks and conditional computation, but the fundamental constraint of sequential computation remains.

Background

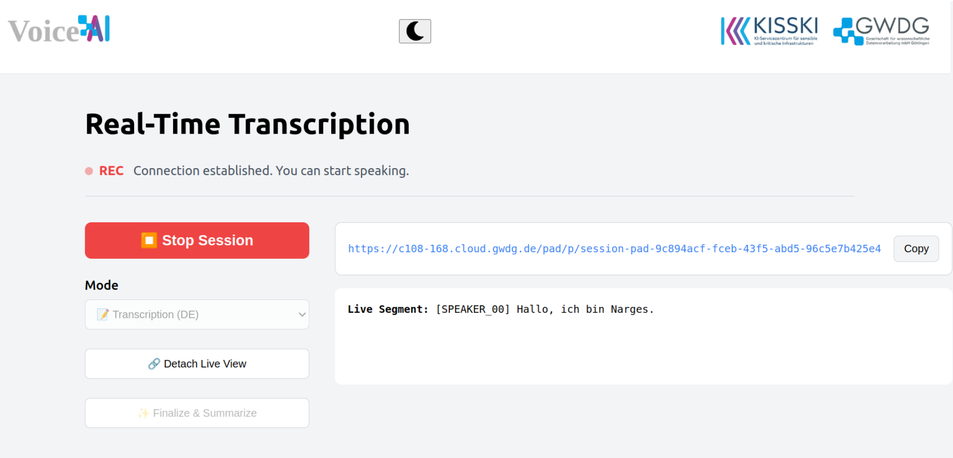
* Attention mechanisms have become integral to sequence modeling and transduction models, allowing for modeling of dependencies without regard to distance in the input or output sequences.
* However, attention mechanisms are often used in conjunction with RNNs.
* Other models, such as the Extended Neural GPU, ByteNet, and ConvS2S, use convolutional neural networks to reduce sequential computation, but still have limitations in relating signals from distant positions.

Proposed Model

* The Transformer model proposed in this work eschews recurrence and relies entirely on an attention mechanism to draw global dependencies between input and output.
* The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for a short time on a large number of GPUs.
* The Transformer uses self-attention, an attention mechanism relating different positions of a single sequence, to compute a representation of the sequence.

Advantages

* Self-attention allows for a constant number of operations to relate signals from two arbitrary input or output positions, making it easier to learn dependencies between distant positions.
* The Transformer's use of self-attention is novel and has not been explored in previous transduction models relying entirely on sequence-aligned RNNs or convolution.



VoiceAI
KISSKI KI-Servicezentrum für sensible und kritische Infrastrukturen
GWDG Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
Stop Session
Detach Live View
Finalize & Summarize
Copy

Real-Time Transcription

● **REC** Connection established. You can start speaking.

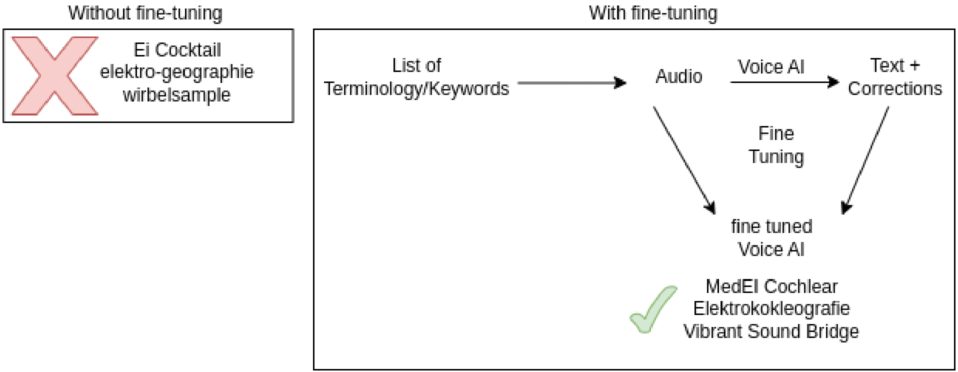
https://c108-168.cloud.gwdg.de/pad/p/session-pad-9c894acf-fceb-43f5-abd5-96c5e7b425e4

Mode

Transcription (DE)

Live Segment: [SPEAKER_00] Hallo, ich bin Narges.

Fine-tuned models:



Without fine-tuning

Ei Cocktail
elektro-geographie
wirbelsample

With fine-tuning

List of Terminology/Keywords → Audio

Audio → Voice AI → Text + Corrections

Fine Tuning

fine tuned Voice AI

MedEl Cochlear
Elektrokokleografie
Vibrant Sound Bridge

Outline

1 Introduction

2 Voice AI-Use Cases

3 **Status**

4 Challenges

5 Roadmap

Status – What Works Today

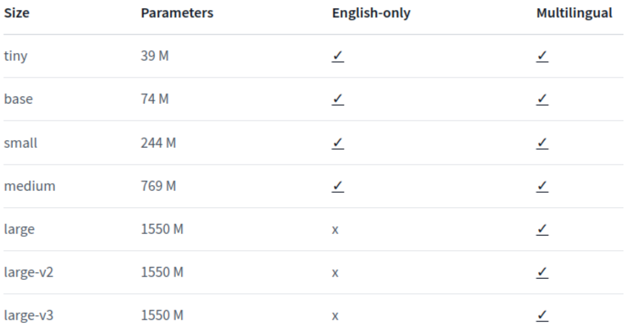
- AI-driven **transcription** and **translation** for 57 languages.
- Accepts multiple file formats: **WAV, MP4, FLAC, etc.**.
- Outputs available in **TXT, SRT, VTT** formats.
  - With SRT & VTT transcription and translation include timestamps
- Used by more than **2,700 active users**.

Status - What works today

- **Accessibility** and ease of use.
  - For individuals who are deaf or hard of hearing
  - Helpful tool for people learning a new language
  - Meetings can be saved and reviewed later, aiding in note-taking and review

Used Model for VoiceAI Service

Whisper is a pre-trained model for automatic speech recognition (ASR) and speech translation. It is trained on 680k hours of labelled data.



Size	Parameters	English-only	Multilingual
tiny	39 M	✓	✓
base	74 M	✓	✓
small	244 M	✓	✓
medium	769 M	✓	✓
large	1550 M	x	✓
large-v2	1550 M	x	✓
large-v3	1550 M	x	✓

Figure: Source: OpenAI, *Whisper large-v3*, 2023

Rollout Feasibility – Can This Be Deployed Widely?

- **Technical Readiness**:
  - Accessible via **API** for integration into applications.
  - Stable performance with minor lags in real-time transcription.
  - **Scalable infrastructure** ready for broader adoption.
- **Adoption Criteria**:
  - Suitable for **academic research, accessibility initiatives, and enterprise usage**.
  - Can be integrated into **educational tools, hybrid learning environments**.

Outline

1 Introduction

2 Voice AI-Use Cases

3 Status

4 Challenges

5 Roadmap

Challenges and Current solution (Implementation Barriers)

1. **Network Speed:**
   - Slow user’s internet may cause timeouts during large file uploads.
   - Suggestion: Use lossless **FLAC** format for efficient upload.
2. **Real-Time Transcriptions/translations:**
   - Previous real-time transcription/translation has a lag of **2-3 seconds**.
   - Current "real" real-time services are **unstable** but under improvement.

Challenges and Current solution

3. **Diarization:**

- Challenges:
  - Adding speaker labels accurately in SRT/VTT formats is challenging.
  - Background noise and overlapping dialogue reduce reliability.
  - Format limitations: Not all tools support custom speaker metadata.
- Possible solutions:
  - Ensemble models combining multiple diarization approaches for higher accuracy Efstathiadis et al. (2025).
  - **Segment-level speaker reassignment** to reduce speaker confusion errors. Boeddeker et al. (2024).
  - Use of **Chain-of-Thought (CoT) prompting** for improved contextual understanding Adedejiet al. (2024).

Challenges and Current Solution

Possible Prompting Scenarios

- Whisper transcribes Chinese audio to **Traditional Chinese** by default.
- Difficulty handling **punctuation correction** (e.g., converting spoken cues like "comma" to actual ‘,‘).

Possible Solutions

- **LLM-based correction**: Refining transcription outputs using AI models.
- **Python libraries/scripts**: Using tools like OpenCC for converting Traditional to Simplified Chinese.
- **Custom rule-based processing**: Implementing prompt manipulation techniques to influence transcription format.

Outline

1 Introduction

2 Voice AI-Use Cases

3 Status

4 Challenges

5 Roadmap

Perspective – How Can We Improve?

- Optimizing real-time transcription for speed and stability.
- Refining diarization for better speaker identification.
- Enhancing accessibility with additional features, such as voice recognition for commands.

References

- OpenAI Whisper Documentation: `https://github.com/openai/whisper`
- FLAC Format Overview: `https://xiph.org/flac/`
- Research on CoT Prompting: `https://arxiv.org/abs/2402.07658v1`
- Research on Segment-Level Speaker Reassignment: `https://arxiv.org/abs/2406.03155`
- Research on LLM-Based Diarization Correction: `https://arxiv.org/html/2406.04927v3`