

15.05.2025

Listen who's whispering

Speech recognition with APEX and Whisper

Dennis Amthor, APEX Connect 2025

code of change



Our key facts



Germany

- Brunswick
- Ratingen
- Hamburg
- Dortmund
- Cologne
- Frankfurt
- Munich
- Berlin

Poland

- Warsaw

Romania

- Cluj-Napoca

Lithuania

- Vilnius
- Kaunas

India

- Pune

850+

Employees

150+

Customers

110+

million €
turnover

Speaker



Dennis Amthor

Consultant APEX/JavaScript

Idea/Motivation

Idea/Motivation

- Voice as data input inside APEX
 - Forms, meeting transcripts, notes
- Controlling APEX functions via voice
 - Searching, filtering
- Feasibility test
 - Complexity
 - Cost
 - Precision
 - Performance
 - Offline capability



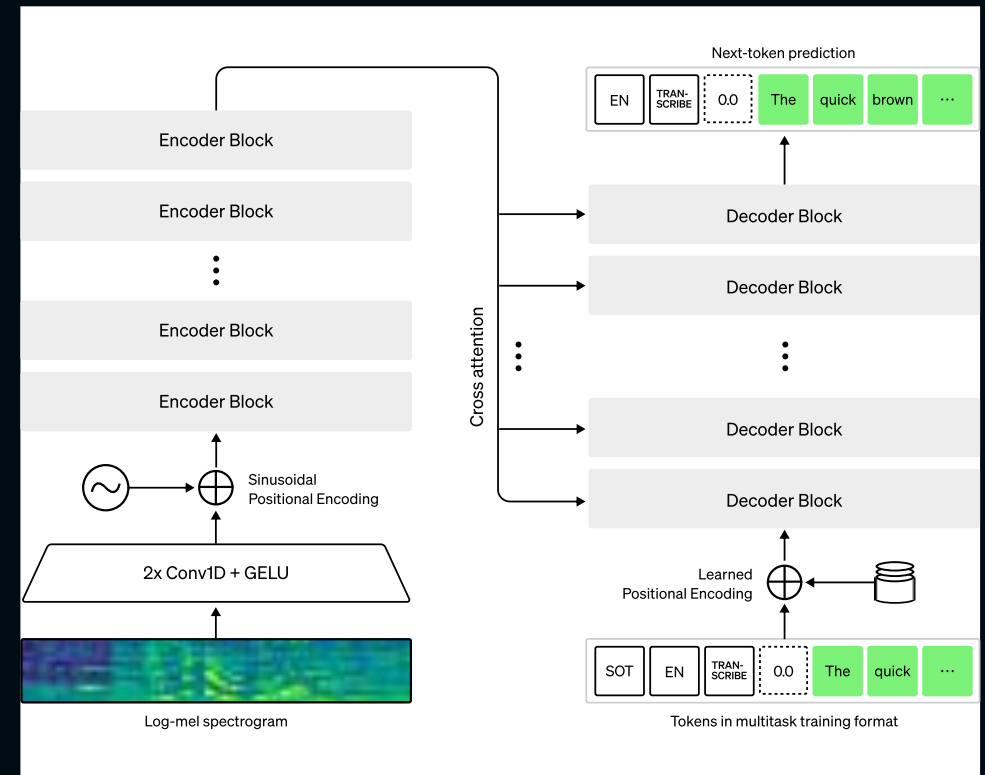
What is Whisper?

What is Whisper?

- Automated Speech Recognition System (ASR) by OpenAI
- Trained with 680,000 hours of audio data (2/3 of which are English)
- Transcription and translation (in English)
- OpenAI: Data source larger and more "diverse" than on other models
 - As a result, less precise overall, but better handling of accents and background noise

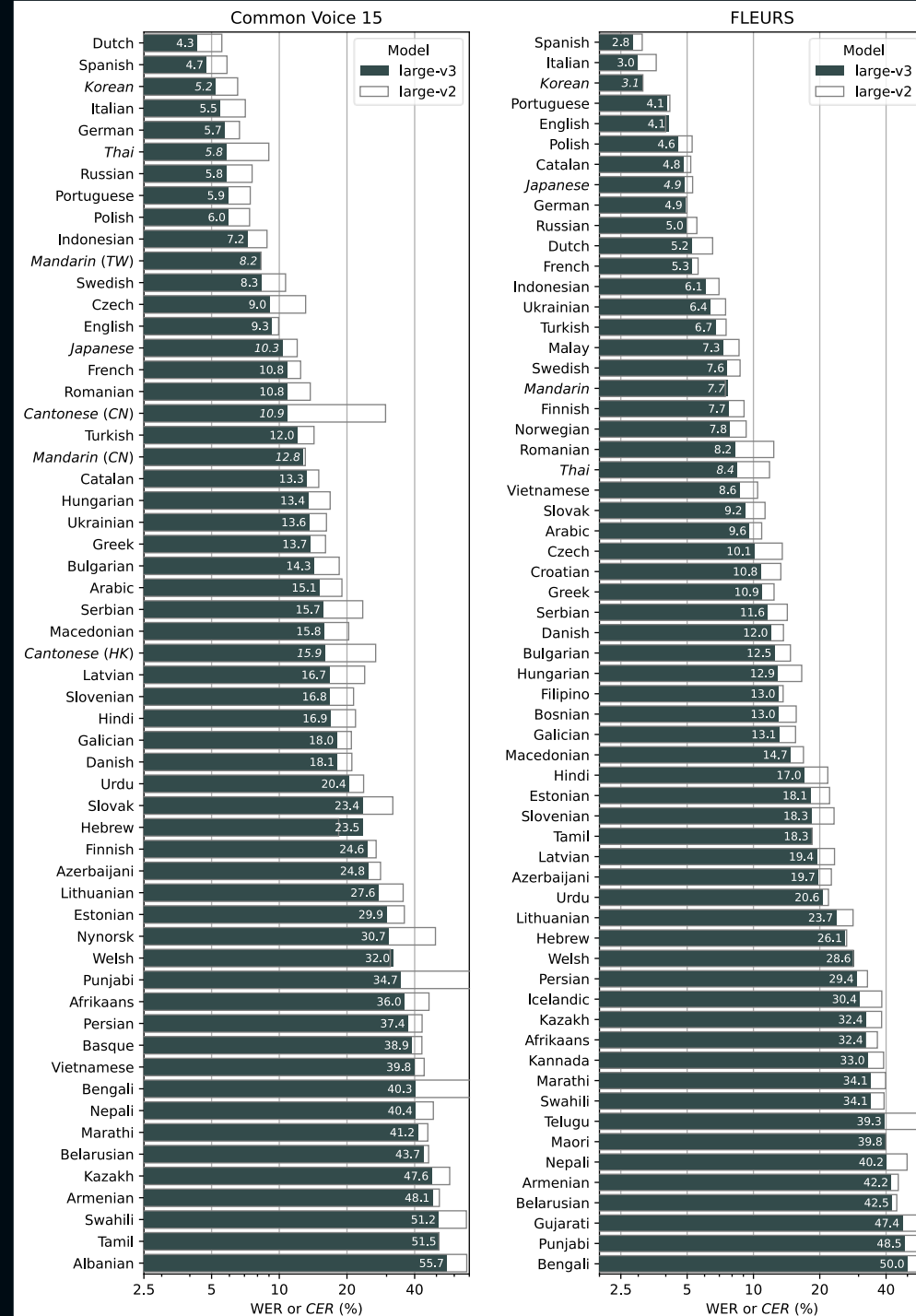
How it works

- Chunking the audio input into blocks of 30 seconds
- Conversion to Mel-Spectrogram
 - Representation of frequencies/time while taking human hearing ability into account
 - Higher resolution in a lower (=audible) frequency range
- Encoder-Decoder-Transformer-Model
 - Audio information as vectors
 - Self- and Cross-Attention
 - Different token types (Text, control, language, timestamps)



Models

- 6 models available
(tiny, base, small, medium, large, turbo)
- Differences in regard to
 - Performance
 - Storage
 - Error rate
- Alternative faster English-only variants available for tiny to medium
- Error rate also highly dependent on the language used



Usage

Whisper

- Calls to OpenAI API
- Model: whisper-1 (\$0.006/minute)
- Alternative: Local integration in Python using PyTorch

Whisper Web

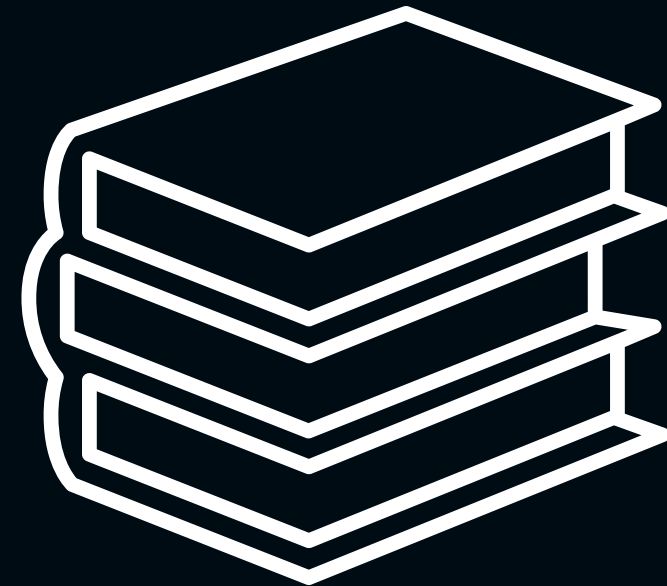
- Web integration using Transformers.js
- Pre-trained models
- Stored in the local cache
- Usage of own models possible after conversion

Transformers

[Huggingface.co](https://huggingface.co)

Transformers (Huggingface.co)

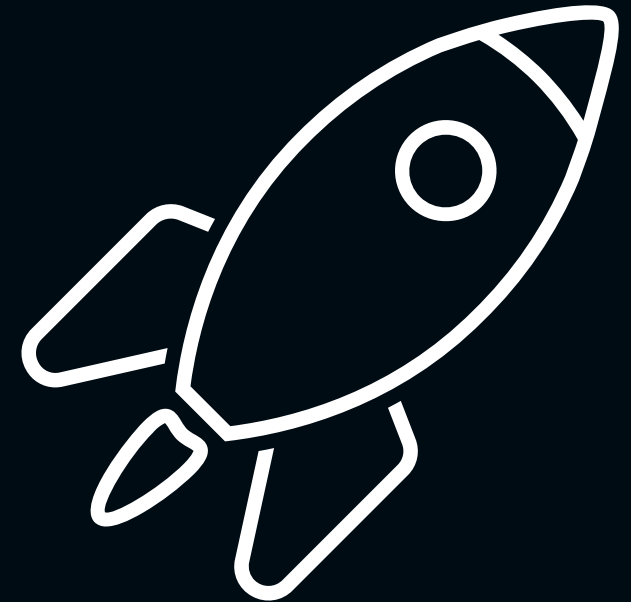
- Library of pre-trained models
 - Text, image, audio, video
 - Various NLP and multimedia tasks
- Models can be used directly and replaced in a modular way (no own training required)
- Unified API and JavaScript integration (Transformers.js)
- Usable directly in the browser (ONNX Runtime)



Implementation

Implementation

- Calls towards Whisper Web/Whisper API using JavaScript (Webpack)
- Integration as APEX plugins
 - Region Plugin for testing/comparison (input methods, technologies)
 - DA plugin for direct use with buttons (examples)
- Whisper Web: Transformer.js with pre-trained model (Xenova/whisper-base)
- Whisper: OpenAI API: AI Service & apex_ai.generate
- Comparison with similar technologies
 - Web Speech API/Speech Recognition
 - (OCI Speech)



Demo

Result

Result

S	
A	
B	
C	

Implementation effort

S	Web Speech API
A	Whisper Web Whisper API
B	
C	

Configuration/Options

S	
A	Whisper Web
B	Whisper API
C	Web Speech API

Precision/Error rate

S	Whisper API
A	Web Speech API
B	Whisper Web
C	

Performance

S	Web Speech API
A	Whisper API
B	
C	Whisper Web

Cost

S	Whisper Web	Web Speech API
A	Whisper API	
B		
C		

Privacy

S	Whisper Web
A	
B	Whisper API
C	Web Speech API

Result

	Whisper Web (base model)	Whisper API	Web Speech API
Implementation effort	A	A	S
Configuration/Options	A	B	C
Precision/Error rate	B	S	A
Performance	C	A	A
Cost	S	A	S
Privacy	S	B	C

Result

	Whisper Web (base model)	Whisper API	Web Speech API
Implementation effort	A	A	OCI Speech
Configuration/Options	A	B	
Precision/Error rate	B	S	?
Performance	C	A	
Cost	S	A	
Privacy	S	B	C

Any questions?