Allgemeine Informationen zu KI

Unsere KI-gestützten Sprachfunktionen basieren auf einem spezialisierten Zusammenspiel mehrerer Modelle. Ein Speech-to-Text-Modell (STT) hört gesprochene Audiodaten und wandelt sie in geschriebenen Text um - das ist der Prozess der Transkription. Ein Text-to-Speech-Modell (TTS) macht das Gegenteil: Es nimmt geschriebenen Text und wandelt ihn in natürlich klingende Sprache um, sodass ein Computer ihn „laut vorlesen“ kann. Ein Large Language Model (LLM) arbeitet ausschließlich mit Text - es versteht, was gesagt wird, und erzeugt sinnvolle schriftliche Antworten. In der Praxis werden diese drei Modelltypen häufig kombiniert: STT wandelt die Sprache einer Person in Text um, das LLM versteht den Inhalt und erstellt eine Antwort, und TTS gibt diese Antwort gesprochen wieder. Damit dies zuverlässig funktioniert, müssen alle diese Modelle mit großen Mengen an Beispieldaten trainiert werden. Die folgenden Abschnitte erläutern, welche Daten wir für unsere Speech-to-Text- und Text-to-Speech-Modelle verwenden.

Speech-to-Text (STT)

Unsere Speech-to-Text-Funktion basiert auf führenden Open-Source-Modellen für Spracherkennung. Konkret nutzen wir OpenAIs Whisper zusammen mit einer eigenen Implementierung mit spezifischen Schutzmechanismen sowie unseren eigenen Fine-Tunings der NVIDIA-Modelle Canary und Parakeet. Indem wir auf diesen starken offenen Grundlagen aufbauen und sie an unsere eigenen Anwendungsfälle anpassen, können wir eine genaue und robuste Transkription über viele Sprachen, Akzente und Aufnahmebedingungen hinweg bereitstellen. Für unser Fine-Tuning und unsere Evaluation verwenden wir die öffentlich verfügbaren, offen lizenzierten Sprachdatensätze, die im Anhang am Ende dieses Dokuments aufgeführt sind. Dort sind jeweils das Quell-Repository, die Lizenz und, sofern vorhanden, die wissenschaftliche Referenz angegeben.

Text-to-Speech (TTS)

Anders als bei unserer Speech-to-Text-Funktion bot die Open-Source-Landschaft im Bereich Text-to-Speech kein Modell, das unseren Qualitätsanforderungen entsprach. Aus diesem Grund haben wir unser eigenes Text-to-Speech-Modell von Grund auf entwickelt. Dafür haben wir es mit den öffentlich verfügbaren, offen lizenzierten Sprachdatensätzen trainiert, die im Anhang am Ende dieses Dokuments aufgeführt sind. Dort sind jeweils das Quell-Repository, die Lizenz und, sofern vorhanden, die wissenschaftliche Referenz angegeben. Der Aufbau eines eigenen Modells gibt uns die volle Kontrolle über Sprachqualität und Aussprache und ermöglicht es uns, aus geschriebenem Text klare, natürlich klingende Sprache zu erzeugen.

Anhang - Herkunft der Trainings- und Evaluationsdaten

Die in diesen Transparenzhinweisen beschriebenen Text-to-Speech- (TTS) und Speech-to-Text-Dienste (STT) wurden mit den unten aufgeführten öffentlich verfügbaren Sprachdatensätzen trainiert und evaluiert. Jeder Eintrag enthält den Namen des Datensatzes, das Hugging-Face-Repository, eine kurze Beschreibung, die Lizenz, unter der der Datensatz bereitgestellt wird, sowie die zugehörige wissenschaftliche Veröffentlichung, sofern vorhanden. Alle Datensätze wurden über Hugging Face bezogen. Die angegebenen Lizenzbedingungen entsprechen dem Stand der Dataset Card oder der begleitenden Dokumentation zum Zeitpunkt der Erstellung; mehrere Datensätze enthalten Namensnennungs- oder Share-Alike-Bedingungen, die bei der Nutzung beachtet werden müssen.

Datensatz	Hugging-Face-Repository	Beschreibung	Lizenz	Referenz / Paper
LEMAS-Dataset	LEMAS-Project/LEMAS-Dataset-train	150.000-Stunden großer, umfangreicher mehrsprachiger Sprachkorpus mit Zeitstempeln auf Wortebene (10 Sprachen).	CC-BY-4.0	LEMAS: A 150K-Hour Large-scale Extensible Multilingual Audio Suite with Generative Speech Models
Omnilingual ASR Corpus	facebook/omnilingual-asr-corpus	Spontane Sprachaufnahmen und Transkriptionen für 348 unterversorgte Sprachen (CC-BY-4.0). Meta FAIR Omnilingual ASR-Projekt.	CC-BY-4.0	Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages (Meta FAIR)
FLEURS	google/fleurs	N-fach paralleler Sprachdatensatz in 102 Sprachen (~12 Stunden pro Sprache), aufgebaut auf FLoRes-101. Teil des XTREME-S-Benchmarks.	CC-BY-4.0	FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech (arXiv:2205.12446)
The People's Speech	MLCommons/peoples_speech	Einer der weltweit größten englischsprachigen Spracherkennungskorpora, lizenziert unter CC-BY-SA / CC-BY 4.0.	CC-BY-4.0 / CC-BY-SA-4.0 (je Clip)	The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset (arXiv:2111.09344)
Emilia	amphion/Emilia-Dataset	Verwendet wurde nur der Emilia-YODAS-Teilbestand (bezogen aus YODAS / YouTube-CC-Audio). Erstellt mit der Open-Source-Pipeline Emilia-Pipe.	CC-BY-4.0 (Emilia-YODAS-Teilbestand)	Emilia: An Extensive, Multilingual, and Diverse Speech Dataset for Large-Scale Speech Generation (arXiv:2407.05361)
Common Voice 17.0	fixie-ai/common_voice_17_0	Mirror von Mozilla Common Voice 17.0, einem massiv mehrsprachigen, crowdsourcing-basierten Korpus transkribierter Sprache (CC0).	CC0-1.0	Common Voice: A Massively-Multilingual Speech Corpus (arXiv:1912.06670, LREC 2020)
LibriTTS-R	mythicinfinity/libritts_r	In der Klangqualität wiederhergestellte Version des mehrsprecherbasierten	CC-BY-4.0	LibriTTS-R: A Restored Multi-

Datensatz	Hugging-Face-Repository	Beschreibung	Lizenz	Referenz / Paper
		englischen TTS-Korpus LibriTTS (aus LibriVox-Hörbüchern).		Speaker Text-to-Speech Corpus (arXiv:2305.18802)
Multilingual LibriSpeech (MLS)	facebook/multilingual_librispeech	Großer mehrsprachiger Korpus aus LibriVox-Hörbüchern, 8 Sprachen (En/De/Nl/Es/Fr/It/Pt/Pl).	CC-BY-4.0	MLS: A Large-Scale Multilingual Dataset for Speech Research (arXiv:2012.03411)
Jenny TTS Dataset	reach-vb/jenny_tts_dataset	~30-stündiger hochwertiger Einsprecher-Datensatz (irisches Englisch), geeignet für TTS-Training.	Custom - Namensnennung erforderlich (Stimme als „Jenny (Dioco)“ nennen)	Keine zugehörige wissenschaftliche Veröffentlichung gefunden.
Thorsten-Voice (TV-44kHz-Full)	Thorsten-Voice/TV-44kHz-Full	~40-stündiger hochwertiger deutscher Einsprecher-Sprachdatensatz (38.000+ transkribierte Aufnahmen von Thorsten Müller).	CC0-1.0	Keine zugehörige wissenschaftliche Veröffentlichung gefunden.
Hi-Fi TTS	MikhailT/hifi-tts	Mehrsprecherbasierter englischer TTS-Datensatz auf Grundlage gemeinfreier LibriVox-Hörbücher und Project-Gutenberg-Texte.	CC-BY-4.0	Hi-Fi Multi-Speaker English TTS Dataset (arXiv:2104.01497)
YODAS-Granary	espnet/yodas-granary	Von YODAS abgeleiteter Teilbestand von NVIDIA Granary, einem mehrsprachigen ASR/AST-Datensatz für 25 europäische Sprachen.	CC-BY-4.0	Granary: Speech Recognition and Translation Dataset in 25 European Languages (arXiv:2505.13404)
FSD50K	Fhrozen/FSD50k	Offener Datensatz mit 51.197 menschlich annotierten Freesound-Audioereignis-Clips (200 AudioSet-Ontologieklassen, ~108 Stunden).	CC-BY-4.0	FSD50K: an Open Dataset of Human-Labeled Sound Events (arXiv:2010.00475)
EuroSpeech	disco-eth/EuroSpeech-24kHz	Norwegischer und schwedischer Teilbestand.	NLOD 2.0 / Riksdag Attribution	EuroSpeech: A Multilingual Speech Corpus