AI General Infos

Our AI-powered voice features rely on a specialized ensemble of models working together. A speech-to-text (STT) model listens to spoken audio and turns it into written text – this is what happens when speech is transcribed. A text-to-speech (TTS) model does the opposite: it takes written text and turns it into natural-sounding spoken audio, so a computer can “read out loud.” A large language model (LLM) works purely with text – it understands what is being said and generates meaningful written responses. In practice, these three are often combined: STT converts a person’s speech into text, the LLM understands and produces a reply, and TTS speaks that reply back. To work well, each of these models has to be trained on large amounts of example data. The sections below explain what data we use for our speech-to-text and text-to-speech models.

Speech-to-Text (STT)

Our speech-to-text capability is built on top of leading open-source speech recognition models. Specifically, we use OpenAI’s Whisper alongside a custom implementation with distinct guardrails and our own fine-tunes of NVIDIA’s Canary and Parakeet models. Building on these strong open foundations and adapting them to our own use cases lets us deliver accurate and robust transcription across many languages, accents, and recording conditions. Our fine-tuning and evaluation use the publicly available, openly licensed speech datasets listed in the appendix at the end of this document, where each dataset’s source repository, license, and academic reference can be found.

Text-to-Speech (TTS)

Unlike our speech-to-text capability, the open-source landscape for text-to-speech did not offer a model that met our quality requirements. For this reason we built our own text-to-speech model from scratch. To do so, we trained it on the publicly available, openly licensed speech datasets listed in the appendix at the end of this document, where each dataset’s source repository, license, and academic reference can be found. Building our own model gives us full control over voice quality and pronunciation and lets us generate clear, natural-sounding speech from written text.

Appendix – Training & Evaluation Data Provenance

The text-to-speech (TTS) and speech-to-text (STT) services described in these transparency notes were trained and evaluated using the publicly available speech datasets listed below. Each entry records the dataset name, its Hugging Face repository, a short description, the license under which it is distributed, and the associated academic publication where one exists. All datasets were sourced from Hugging Face. License terms shown reflect the dataset card or accompanying documentation at the time of writing; several carry attribution or share-alike conditions that must be observed in use.

Dataset	Hugging Face repository	Description	Licence	Reference / paper
LEMAS-Dataset	LEMAS-Project/LEMAS-Dataset-train	150K-hour large-scale multilingual speech corpus with word-level timestamps (10 languages).	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	Spontaneous speech recordings and transcriptions for 348 under-served languages (CC-BY-4.0). Meta FAIR Omnilingual ASR project.	CC-BY-4.0	Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages (Meta FAIR)
FLEURS	google/fleurs	N-way parallel speech dataset in 102 languages (~12 hrs/language), built on FLoRes-101. Part of the XTREME-S benchmark.	CC-BY-4.0	FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech (arXiv:2205.12446)
The People's Speech	MLCommons/peoples_speech	One of the world's largest English speech recognition corpora, licensed CC-BY-SA / CC-BY 4.0.	CC-BY-4.0 / CC-BY-SA-4.0 (per clip)	The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset (arXiv:2111.09344)
Emilia	amphion/Emilia-Dataset	Only the Emilia-YODAS subset was used (sourced from YODAS / YouTube CC audio). Built with the open-source Emilia-Pipe pipeline.	CC-BY-4.0 (Emilia-YODAS subset)	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 of Mozilla Common Voice 17.0, a massively-multilingual crowdsourced corpus of transcribed speech (CC0).	CC0-1.0	Common Voice: A Massively-Multilingual Speech Corpus (arXiv:1912.06670, LREC 2020)
LibriTTS-R	mythicinfinity/libritts_r	Sound-quality-restored version of the LibriTTS multi-speaker English TTS corpus (from LibriVox audiobooks).	CC-BY-4.0	LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus (arXiv:2305.18802)

Dataset	Hugging Face repository	Description	Licence	Reference / paper
Multilingual LibriSpeech (MLS)	facebook/multilingual_librispeech	Large multilingual corpus from LibriVox audiobooks, 8 languages (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-hour high-quality single-speaker (Irish English) voice dataset suitable for TTS training.	Custom — attribution required (credit voice as "Jenny (Dioco)")	*No associated academic paper found.*
Thorsten-Voice (TV-44kHz-Full)	Thorsten-Voice/TV-44kHz-Full	~40-hour high-quality single-speaker German speech dataset (38,000+ transcribed recordings by Thorsten Müller).	CC0-1.0	*No associated academic paper found.*
Hi-Fi TTS	MikhailT/hifi-tts	Multi-speaker English TTS dataset based on LibriVox public-domain audiobooks and Project Gutenberg texts.	CC-BY-4.0	Hi-Fi Multi-Speaker English TTS Dataset (arXiv:2104.01497)
YODAS-Granary	espnet/yodas-granary	YODAS-derived subset of NVIDIA Granary, a multilingual ASR/AST dataset covering 25 European languages.	CC-BY-4.0	Granary: Speech Recognition and Translation Dataset in 25 European Languages (arXiv:2505.13404)
FSD50K	Fhrozen/FSD50k	Open dataset of 51,197 human-labeled Freesound sound-event clips (200 AudioSet Ontology classes, ~108 hours).	CC-BY-4.0	FSD50K: an Open Dataset of Human-Labeled Sound Events (arXiv:2010.00475)
EuroSpeech	disco-eth/EuroSpeech-24kHz	Norwegian & Swedish Subset	NLOD 2.0 / Riksdag Attribution	EuroSpeech: A Multilingual Speech Corpus