

STT LV Datasets

Dataset	Info	Storage path (asya-5)	Number of hours	Status
Youtube_unlabeled	Unlabeled scraped speech data from youtube.	/media/storage_2/data/raw/speech_audio/data_stt_lv_yt_unlabeled_all	~1800h + pp_unlabeled hours.	Ready
Pitchpatterns unlabeled speech data.	Unlabeled conversation speech data fro pretraining.	/media/storage_2/data/raw/speech_audio/data_stt_lv_pp	~1000h	Processing
Youtube_labeled	Labeled scraped speech data from youtube with transcripts.	asya_5: /media/storage_2/data/raw/speech_audio/data_stt_lv_yt_labeled/audio	~130h raw labeled noisy.	Ready
Movies dataset	Scrapped movies with subtitles.	asya_5: /media/storage_2/data/raw/speech_audio/data_stt_lv_yt_labeled/audio_movies	42h	Ready
Google fleurs dataset / latvian part.	Part of FLEURS multilingual dataset.	asya_5: /media/storage_2/data/raw/speech_audio/data_stt_lv_fleurs	10.5h	Ready
Mozilla commonvoice latvian.	Mozilla labeled latvian stt data.	asya_5: /media/storage_2/data/raw/speech_audio/data_stt_lv_cv	~7h	Ready
Pitchpatterns labeled speech data.	Labeled conversation speech data for finetuning/validation?	/media/storage_2/data/raw/speech_audio/data_stt_lv_pp/audio	~40min	Ready

Text datasets:

Dataset	Info	Storage path (asya-5)	Words count	Status
lv_wikipedia	Scrapped wikipedia articles in latvian.	/media/storage_2/data/raw/lv_text_corpus/lv_wikipedia	-	Ready
Books	Processed text data from available latvian books.	/media/storage_2/data/raw/lv_text_corpus/data_lv_books	~58mil	Ready
Research papers/ thesis etc	Latvian papers/ bachelors/masters/ thesis...	-	-	Processing